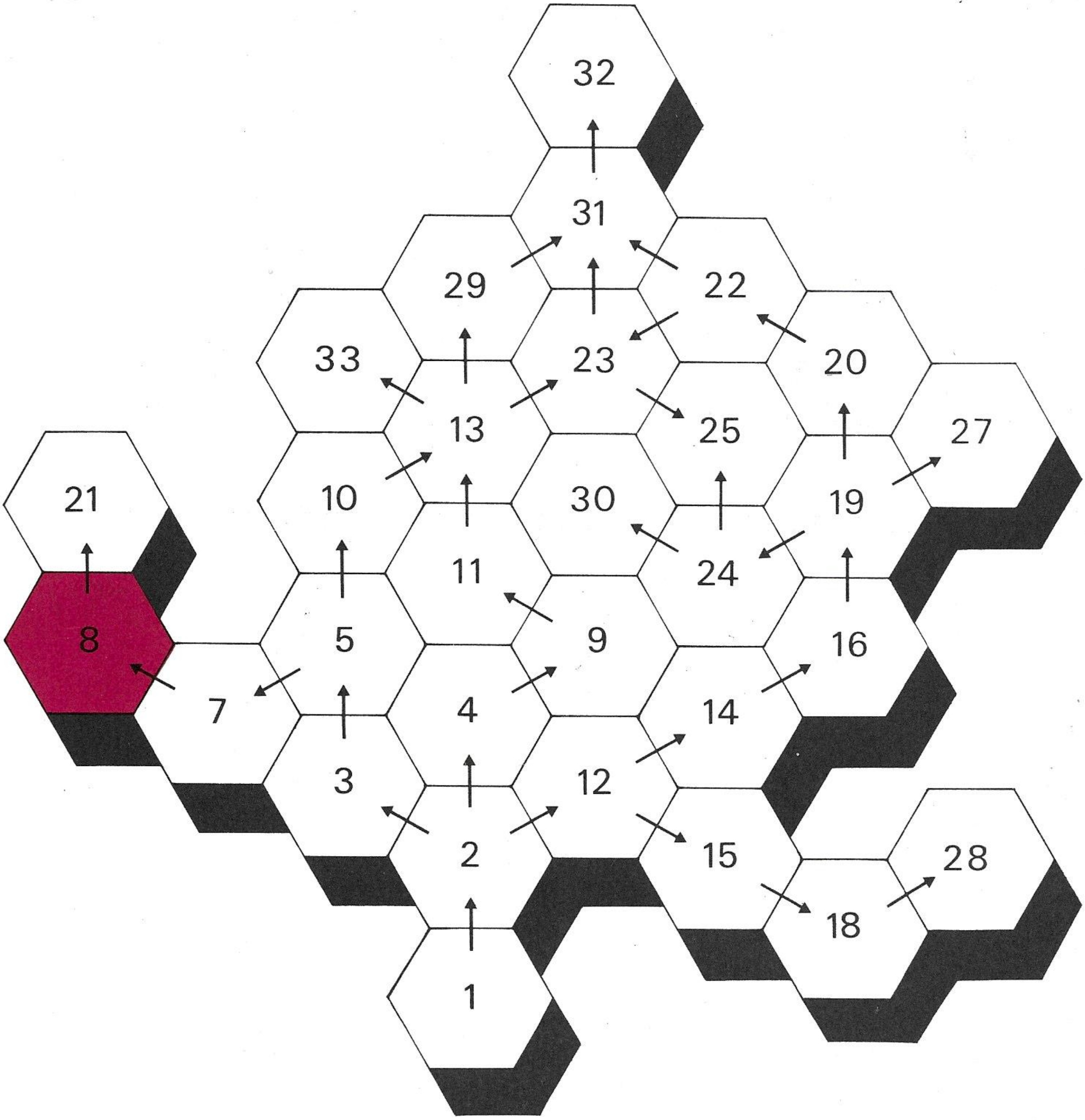
M201 8

THE OPEN UNIVERSITY

Mathematics: A Second Level Course

Linear Mathematics Unit 8

Numerical Solution of Simultaneous Algebraic Equations



The Open University

Mathematics: A Second Level Course

Linear Mathematics   Unit 8

NUMERICAL SOLUTION OF SIMULTANEOUS ALGEBRAIC EQUATIONS

Prepared by the Linear Mathematics Course Team

The Open University Press

The Open University Press Walton Hall Milton Keynes MK7 6AA

First published 1972. Reprinted 1976

Designed by the Media Development Group of the Open University

Printed in Great Britain by
Martin Cadbury

SBN 335 01097 0

This text is one in a series of units that make up the correspondence element of an Open University Second Level Course. The complete list of units in the course is given at the end of this text.

For general availability of supporting material referred to in this text, please write to the Director of Marketing, The Open University, P.O. Box 81, Walton Hall, Milton Keynes, MK7 6AT.

Further information on Open University courses may be obtained from the Admissions Office, The Open University, P.O. Box 48, Walton Hall, Milton Keynes, MK7 6AB.

1.2

Contents

Set Books

D. L. Kreider, R. G. Kuller, D. R. Ostberg and F. W. Perkins, *An Introduction to Linear Analysis* (Addison-Wesley, 1966).

E. D. Nering, *Linear Algebra and Matrix Theory* (John Wiley, 1970).

It is essential to have these books; the course is based on them and will not make sense without them.

Conventions

Before working through this correspondence text make sure you have read *A Guide to the Linear Mathematics Course.* Of the typographical conventions given in the Guide the following are the most important.

The set books are referred to as:

K for *An Introduction to Linear Analysis*
N for *Linear Algebra and Matrix Theory*

All starred items in the summaries are examinable.

References to the Open University Mathematics Foundation Course Units (The Open University Press, 1971) take the form *Unit M100 3, Operations and Morphisms.*

Note

Please note that this text is not based on the set books for the course.

8.0 INTRODUCTION

In *Unit 7, Recurrence Relations*[*] we gave an introduction to the things we have to consider when we try to find numerical answers to general problems, and illustrated these considerations, and their analysis, with respect to computations involving linear recurrence relations.

In this unit we look from the same point of view at an even more important and ubiquitous problem, the solution of simultaneous linear algebraic equations. As before, we are concerned with two main aspects of stability:

(i) Can the *problem* exhibit *inherent instability* (ill-conditioning), and how would we recognize this defect?
(ii) Can some *methods* exhibit *induced instability*, and can we find methods which avoid it?

These topics have already had some previous introductory attention. For example, in *Unit M100 28, Linear Algebra IV* we looked at the possibilities of ill-conditioning, observing that for the case of two equations the ill-conditioning is related to the difficulty of finding accurately the intersection of two nearly parallel lines in a plane. We also noticed that the ill-conditioning in general is related to the near-singularity of the matrix of the equations, and indeed recommended the computation of the inverse with a view to revealing the possibility of inherent instability.

We did not investigate the *stability of methods* in *Unit M100 28*, but we did give some measure of the relative efficiency of a few methods in relation to the amount of arithmetic involved. Suppose we have a linear problem in the form

$$A\mathbf{x} = \mathbf{b},$$

where A is a real square matrix of order n, $\mathbf{b}$ a known one-column matrix representing a vector in R^n, and $\mathbf{x}$ the one-column matrix representing the required solution vector with elements $x_1, x_2, \ldots, x_n$. There is a formal "mathematical" solution known as Cramer's rule, given by

$$x_r = \frac{\Delta_r}{\Delta}, \qquad r = 1, 2, \ldots, n,$$

where Δ is the determinant of the matrix A, and Δ_r is the determinant of the matrix which is obtained by replacing the rth column of A by the column $\mathbf{b}$. This method we rightly cast out because the amount of work involved is prohibitively large.

On the other hand, a process of elimination, invented by Gauss and closely related to the calculation of the Hermite normal form (see *Unit 3, Hermite Normal Form*), turned out to be an efficient method, at least with respect to the volume of computation involved.

In this unit we take up these questions in more detail. Following the pattern of *Unit 7, Recurrence Relations*, we first give, in Section 8.1, some examples of ill-conditioning for a physical problem, in which at least some of the data are uncertain (due perhaps to inaccurate measurement or ignorance), and for a mathematical problem in which the data, though exact, cannot be stored exactly in our computing machine. Since ill-conditioning (inherent instability) is a function only of the problem, all the examples at this stage are treated with exact arithmetic, so that the ill-conditioning is separated from the (induced instability) effects of rounding errors in the computation.

In sub-section 8.1.3, we emphasize the fact that ill-conditioning is a property of the problem, not the method, by illustrating the curious fact that

[*] When referring to *Unit 7*, we quote the short version of its title only.

while the solutions $x_1, x_2, \ldots, x_n$ of the linear equations may be uncertain by relatively large amounts in respect of small changes (uncertainties) in the data, it might well happen that some not very different problem may be relatively well-conditioned. Such a problem, for example, which arises quite frequently in practice, is the evaluation of the single number

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n,$$

$(x_1, x_2, \ldots, x_n)$ being the solution vector of $A\mathbf{x} = \mathbf{b}$, for specified numbers $c_1, c_2, \ldots, c_n$. For some cs, the number y can be quite well determined, because the possibly large uncertainties in the xs are not independent.

In Section 8.2 we examine the most popular and useful direct method, the Gauss elimination method, and its variant attributed to Jordan, and we note their relation with our work in *Unit 3* on the production of the Hermite normal form. With exact arithmetic this gives the unique exact solution if A is non-singular, or reveals any singularity if it exists. We also show (in sub-section 8.2.4) that the method can be expressed in terms of an important and fundamental matrix theorem concerned with the expression of a matrix as the product of two simpler (triangular) matrices.

Having established the method, we then examine its performance, investigating whether or not it can exhibit induced instability when we have to use inexact computer arithmetic. Sub-section 8.3.1 discusses an example in which induced instability occurs, and the reason for this is analysed in general terms in sub-section 8.3.2. The analysis is constructive, in that it shows the possibility of making small changes in the basic method which virtually eliminate the induced instability. This stable variant is described in sub-section 8.3.3.

The error analysis, called *backward error analysis*, which we use in the analysis described above, unfortunately does not give any bound for the error of our computed solution, so that we have to discuss this separately. In Section 8.4, we show how to correct an approximate solution with a small amount of extra work, and this itself answers the question about the accuracy of the first approximation. It also partially answers the question, so far unexplored, of how we most easily detect any ill-conditioning in the original problem.

The methods discussed in this unit are called *direct methods*, since if everything is "exact" the solutions are obtained exactly in a finite number of operations.

In another class of methods, called *iterative methods*, we never get the exact solutions, but approach them more and more closely in a sequence of operations, which is terminated when our approximation is sufficiently accurate in some particular context. Iterative methods were introduced very briefly in *Unit M100 28*. We do not include them here because, though they are very important, their main application is in the treatment of matrices of rather special types, for example, those arising in the numerical solution of partial differential equations.

8.1 ILL-CONDITIONING OF THE LINEAR EQUATION PROBLEM

8.1.1 A Physical Problem

As usual in our numerical work, we look for possibilities of ill-conditioning in any given problem. We begin by illustrating the possibility with a simple example. Suppose we have the *physical* problem

$$\begin{aligned} \tfrac{1}{2}x_1 + \tfrac{1}{3}x_2 + \tfrac{1}{4}x_3 &= b_1 + e_1 \\ \tfrac{1}{3}x_1 + \tfrac{1}{4}x_2 + \tfrac{1}{5}x_3 &= b_2 + e_2 \\ \tfrac{1}{4}x_1 + \tfrac{1}{5}x_2 + \tfrac{1}{6}x_3 &= b_3 + e_3 \end{aligned}$$

in which the coefficients are known exactly, whereas the numbers b_r are subject to measurement errors of amounts e_r, and suppose $|e_r| \leqslant 10^{-3}$ for $r = 1, 2, 3$.

In terms of matrices, we may write our physical problem as

$$A\mathbf{x} = \mathbf{b} + \mathbf{e}$$

and whatever the values of the b_r, the contribution to the solution from the uncertainties e_r is $A^{-1}\mathbf{e}$.

A simple computation (see *Unit M100 26, Linear Algebra III*) gives

$$A^{-1} = \begin{bmatrix} 72 & -240 & 180 \\ -240 & 900 & -720 \\ 180 & -720 & 600 \end{bmatrix}$$

By inspecting the signs of the elements of A^{-1}, we see that the maximum effect of the e_r occurs when $e_1 = \pm 10^{-3}$, $e_2 = \mp 10^{-3}$ and $e_3 = \pm 10^{-3}$, and that the solution $\mathbf{x}$ has possible uncertainties of amounts ± 0.492, ∓ 1.860, ± 1.500.

These uncertainties are very much larger than those of the data, and we have a severe case of *absolute* inherent instability (ill-conditioning). Does this matter? What is the *relative* effect, which may not be serious if the "true" solution $\mathbf{x} = A^{-1}\mathbf{b}$ is also large?

Well, if $b_1 = 1$, $b_2 = -1$, $b_3 = 1$, then

$$x_1 = 492,\ x_2 = -1860,\ x_3 = 1500.$$

The *relative* uncertainty in the answer is precisely the same as that of the data. So with $b_1 = 1$, $b_2 = -1$, $b_3 = 1$, this problem is not ill-conditioned in a relative sense.

Unfortunately there are some right-hand sides, of comparable magnitude with those of the previous example, for which the true solution $\mathbf{x}$ is quite small. For example, with $b_1 = 0.95$, $b_2 = 0.67$, $b_3 = 0.52$, we find

$$x_1 = 1.2,\ x_2 = 0.6,\ x_3 = 0.6.$$

With the same uncertainty $|e_r| \leqslant 10^{-3}$ in the right-hand sides, all we can say about our solution is that

$$\begin{aligned} &x_1 = 1.20 \pm 0.492,\ x_2 = 0.60 \mp 1.860, \\ &x_3 = 0.60 \pm 1.500. \end{aligned}$$

Here we have a very ill-conditioned situation. Not a single significant figure in the answer is "meaningful", that is, can be quoted as being certainly correct for all uncertainties in the data within the known limits!

Exercises

1. Consider the matrix

$$A = \begin{bmatrix} \frac{1}{5} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{7} \end{bmatrix}$$

and suppose that we want to solve the equations

$$A\mathbf{x} = \mathbf{b} + \mathbf{e},$$

where

$$|e_r| \leqslant 0.5 \times 10^{-2}.$$

Show that the problem is absolutely ill-conditioned.

2. Find a vector **b** for which the problem of Exercise 1 is relatively well-conditioned, and another vector **b**, with elements of much the same "size", for which the problem is relatively ill-conditioned.

Solutions

1. We find

$$A^{-1} = \begin{bmatrix} 180 & -210 \\ -210 & 252 \end{bmatrix},$$

and the maximum uncertainty in the solution, which is $A^{-1}\mathbf{e}$, occurs when $e_1 = \pm 0.5 \times 10^{-2}$, $e_2 = \mp 0.5 \times 10^{-2}$, that is, the uncertainties in the data e_1 and e_2 have their maximum values with different signs. We find $A^{-1}\mathbf{e}$, with these values for e_1 and e_2, has elements ± 1.95 and ∓ 2.31. These are the uncertainties in the respective solutions x_1 and x_2. They are very much larger than e_1 and e_2, and the problem is, therefore, absolutely ill-conditioned.

2. An example is given by $b_1 = 1$, $b_2 = -1$, for then the values of x_1 and x_2 are 390 and -462, and the maximum values of the uncertainties in x_1 and x_2 are only 0.5% of the values of x_1 and x_2. This is the same ratio as that of the e_r to the b_r, and the problem is relatively well-conditioned.

Relative ill-conditioning occurs when $A^{-1}\mathbf{e}$ has relatively large importance. We then want b_1 and b_2 to be such that both $180\, b_1 - 210\, b_2$ and $-210\, b_1 + 252\, b_2$ are small for values of b_1 and b_2 of the order of unity. Inspection shows that $b_1 = 1$, $b_2 = \frac{6}{7}$ is one such pair, which gives

$$x_1 = 0 \pm 1.95,\ x_2 = 6.0 \mp 2.31,$$

with severe relative ill-conditioning.

A little further experiment gives, for $b_1 = 1$, $b_2 = \frac{71}{84}$, for example, the results

$$x_1 = 2.5 \pm 1.95,\ x_2 = 3.0 \mp 2.31,$$

and the ill-conditioning is here even more pronounced.

8.1.2 A Mathematical Problem

Consider now a similar example of a *mathematical* problem, in which all the data are exact but not capable of exact storage. We know from previous work that if the physical problem is ill-conditioned, then it is probably difficult to get an accurate solution (which here *is* meaningful) to the mathematical problem. Suppose that in the problem posed in the first paragraph of sub-section 8.1.1 the right-hand sides are exactly 0.95, 0.67 and 0.52, so that the exact solution is $x_1 = 1.2$, $x_2 = 0.6$, $x_3 = 0.6$, but that we have only a three-digit machine so that the left-hand coefficients cannot be stored exactly. We then effectively solve the slightly "perturbed" problem

$$\begin{aligned} 0.500x_1 + 0.333x_2 + 0.250x_3 &= 0.95 \\ 0.333x_1 + 0.250x_2 + 0.200x_3 &= 0.67 \\ 0.250x_1 + 0.200x_2 + 0.167x_3 &= 0.52 \end{aligned}$$

The solution of this problem, avoiding any further arithmetic errors (which our poor old machine would almost certainly make) is

$$x_1 = 1.114, \; x_2 = 0.937, \; x_3 = 0.324,$$

correctly rounded to three decimal places. It is far from the "truth", and we should clearly need to store the *data* far more accurately to produce a tolerably accurate solution.

These results confirm the expectations stimulated by our previous experience in *Unit 7, Recurrence Relations.*

We repeat that what we have illustrated is *inherent instability* or ill-conditioning; it is a function only of the problem. The only errors were "physical" uncertainty in the data or "mathematical" errors in the storage of exact data. The method of solution was quite immaterial, and in fact we used *exact* methods on *inexact* data. Induced instability was nowhere in evidence.

Although we have indicated that ill-conditioning may exist we have not indicated how we might detect or analyse this situation. This we defer until sub-section 8.4.2, after we have found a stable method of solution.

8.1.3 Effect of Correlated Uncertainties in the Solution

Another important point, not made explicitly in *Unit 7*, is that it is nonsense to talk of a recurrence relation as being ill-conditioned. It all depends on what we are trying to do with it. In *Unit 7*, for example, we observed that the *problem* defined by the recurrence relation

$$y_{r+1} = 1 - ry_r,$$

and associated condition $y_0 = 1 - e^{-1}$, is certainly ill-conditioned for the determination of $y_1, y_2, \ldots,$ but the same recurrence relation with condition $y_n = 0$ is perfectly well-conditioned for the determination of y_{n-1}, $y_{n-2}, \ldots, y_1, y_0$.

In a similar way we cannot talk of an ill-conditioned matrix, except in particular contexts, such as the solution of a *particular* set of linear equations, the inversion of the matrix, or the determination of its eigenvalues and eigenvectors.

It can even happen that the solution of a particular set of linear equations is ill-conditioned, whereas some linear combination of the components of the solution is reasonably well-conditioned, and this may be the number we are trying to find. The point here is that the errors in the individual components of the solution are not independent, and *some* linear combination may cause some cancellation of the errors.

Consider, for example, the problem in sub-section 8.1.1 of finding the solution of the equations

$$\begin{aligned}
\tfrac{1}{2}x_1 + \tfrac{1}{3}x_2 + \tfrac{1}{4}x_3 &= 0.95 + e_1\\
\tfrac{1}{3}x_1 + \tfrac{1}{4}x_2 + \tfrac{1}{5}x_3 &= 0.67 + e_2\\
\tfrac{1}{4}x_1 + \tfrac{1}{5}x_2 + \tfrac{1}{6}x_3 &= 0.52 + e_3
\end{aligned}$$

where $|e_r| \leqslant 10^{-3}$, and where we obtained the extremely ill-conditioned results

$$x_1 = 1.20 \pm 0.492,\ x_2 = 0.60 \mp 1.860,$$
$$x_3 = 0.60 \pm \text{i}.500.$$

Suppose, however, that our problem is not the computation of the x_r but the determination of a quantity

$$y = c_1x_1 + c_2x_2 + c_3x_3,$$

for some given numbers c_r. There are, in fact, some numbers c_r for which *this* problem is *not* badly conditioned. Let us try to find some.

The vector $\mathbf{x}$ of the solution is

$$\begin{aligned}
\mathbf{x} = A^{-1}(\mathbf{b} + \mathbf{e}) &= \begin{bmatrix} 72 & -240 & 180\\ -240 & 900 & -720\\ 180 & -720 & 600 \end{bmatrix} \begin{bmatrix} 0.95 + e_1\\ 0.67 + e_2\\ 0.52 + e_3 \end{bmatrix}\\
&= \begin{bmatrix} 1.20\\ 0.60\\ 0.60 \end{bmatrix} + \begin{bmatrix} 72 & -240 & 180\\ -240 & 900 & -720\\ 180 & -720 & 600 \end{bmatrix} \begin{bmatrix} e_1\\ e_2\\ e_3 \end{bmatrix}
\end{aligned}$$

The required value of y is then

$$\begin{aligned}
y &= [c_1 \quad c_2 \quad c_3] \left\{ \begin{bmatrix} 1.20\\ 0.60\\ 0.60 \end{bmatrix} + \begin{bmatrix} 72 & -240 & 180\\ -240 & 900 & -720\\ 180 & -720 & 600 \end{bmatrix} \begin{bmatrix} e_1\\ e_2\\ e_3 \end{bmatrix} \right\}\\
&= 1.20c_1 + 0.60c_2 + 0.60c_3 + \mathbf{de},
\end{aligned}$$

where the row-vector $\mathbf{d} = [d_1 \ d_2 \ d_3]$ has components

$$\begin{aligned}
d_1 &= 72c_1 - 240c_2 + 180c_3\\
d_2 &= -240c_1 + 900c_2 - 720c_3\\
d_3 &= 180c_1 - 720c_2 + 600c_3.
\end{aligned}$$

All we require is that d_1, d_2, and d_3 should be small for "reasonable" values of the c_r, even though A^{-1} has large elements. By inspection we see that $c_1 = 1$, $c_2 = 1$, $c_3 = 1$ will give d_r which are quite small relative to the elements of A^{-1}. In fact, $d_1 = +12$, $d_2 = -60$, $d_3 = +60$, and the maximum possible value of $|\mathbf{de}|$, with $|e_r| \leqslant 10^{-3}$, is $132 \times 10^{-3} = 0.132$.

We have therefore produced the result

$$y = 2.400 \pm 0.132,$$

and the statement $y = 2.4$ can be made with the confident assertion that its maximum possible relative uncertainty is less than 6%, a quite satisfactory result in many physical problems.

Exercise

Consider the equations

$$\tfrac{1}{5}x_1 + \tfrac{1}{6}x_2 = 1 + e_1$$
$$\tfrac{1}{6}x_1 + \tfrac{1}{7}x_2 = \tfrac{71}{84} + e_2$$

where $|e_1| \leqslant 0.5 \times 10^{-2}$, $|e_2| \leqslant 0.5 \times 10^{-2}$.

We saw in Exercise 2 of sub-section 8.1.1 that all we can say about the solution is

$$x_1 = 2.5 \pm 1.95, \; x_2 = 3.0 \mp 2.31,$$

a very ill-conditioned situation with not a single meaningful figure.

Suppose, however, that our problem requires the value not of x_1 and x_2 but of $y = x_1 + 0.8x_2$. Show that

$$y = 4.9 \pm 0.102,$$

a reasonably well-conditioned situation.

Solution

From the computation of the inverse in Exercise 1 of sub-section 8.1.1, we have

$$\mathbf{x} = \begin{bmatrix} 180 & -210 \\ -210 & 252 \end{bmatrix} \begin{bmatrix} 1 + e_1 \\ \tfrac{71}{84} + e_2 \end{bmatrix}$$
$$= \begin{bmatrix} 2.5 \\ 3.0 \end{bmatrix} + \begin{bmatrix} 180e_1 - 210e_2 \\ -210e_1 + 252e_2 \end{bmatrix}$$
$$y = [1 \quad 0.8] \begin{bmatrix} 2.5 \\ 3.0 \end{bmatrix} + [1 \quad 0.8] \begin{bmatrix} 180e_1 - 210e_2 \\ -210e_1 + 252e_2 \end{bmatrix}$$
$$= 4.9 + 12e_1 - 8.4e_2.$$

The worst contribution from the uncertainties is

$$\pm [12(0.5 \times 10^{-2}) - 8.4(-0.5 \times 10^{-2})] = \pm 0.102.$$

8.1.4 Summary of Section 8.1

In this section and the introduction we defined the terms

direct method	(page **C** 6)	* * *
iterative method	(page **C** 6)	* * *
physical problem	(page **C** 7)	* *
mathematical problem	(page **C** 9)	* *

Techniques

1. Solve $A\mathbf{x} = \mathbf{b}$, given specified errors in A or $\mathbf{b}$ and hence say whether or not the problem is ill-conditioned. *
2. If such a problem is ill-conditioned, determine whether the problem of finding $\sum_{s=1}^{n} c_s x_s$, given $c_1, c_2, \ldots, c_n$, is well-conditioned. *

8.2 DIRECT METHODS

8.2.1 Gauss Elimination

Turning now to a direct method for solving simultaneous equations, we shall first describe the *elimination method of Gauss* in its basic form. For simplicity, we deal in detail with three or four equations in three or four unknowns, considering $A\mathbf{x} = \mathbf{b}$ but recording only the elements of A and $\mathbf{b}$ in an array like:

$$\begin{array}{ccccc} & A & & & \mathbf{b} \\ a_{11} & a_{12} & a_{13} & & b_1 \\ a_{21} & a_{22} & a_{23} & & b_2 \\ a_{31} & a_{32} & a_{33} & & b_3 \end{array}$$

The method has two essential parts.

(i) *Elimination*

Evaluate $m_{21} = -a_{21}/a_{11}$, and add m_{21} times the first row of the array to the second row to produce a new second row with first element zero. This is possible if $a_{11} \neq 0$.

Evaluate $m_{31} = -a_{31}/a_{11}$, and perform the corresponding operation with the third row. Again, this is possible if $a_{11} \neq 0$.

The resulting array looks like:

$$\begin{array}{ccccc} & A^{(2)} & & & \mathbf{b}^{(2)} \\ a_{11} & a_{12} & a_{13} & & b_1 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & & b_2^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & & b_3^{(2)} \end{array}$$

where $a_{22}^{(2)} = a_{22} + m_{21}\, a_{12}$, etc.

Now ignore the first row, form $m_{32} = -a_{32}^{(2)}/a_{22}^{(2)}$, and add m_{32} times the second row to the third to eliminate the element $a_{32}^{(2)}$. This is possible if $a_{22}^{(2)} \neq 0$. The resulting and final array looks like:

$$\begin{array}{ccccc} & A^{(3)} = U & & & \mathbf{b}^{(3)} = \mathbf{c} \\ a_{11} & a_{12} & a_{13} & & b_1 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & & b_3^{(3)} \end{array}$$

Here U is an *upper triangular matrix*,* and we have "reduced" the equations $A\mathbf{x} = \mathbf{b}$ to the equivalent form $U\mathbf{x} = \mathbf{c}$.

This is the "elimination" part of the algorithm, and the general method for n equations should be clear from this description. The required solution is obtained in the second part of the algorithm, the so-called "back-substitution".

(ii) *Back-substitution*

Compute $x_3 = b_3^{(3)}/a_{33}^{(3)}$, which is possible if $a_{33}^{(3)} \neq 0$. Substitute for x_3 in the second row of $U\mathbf{x} = \mathbf{c}$ and compute

$$x_2 = (b_2^{(2)} - a_{23}^{(2)} x_3)/a_{22}^{(2)},$$

which is possible if $a_{22}^{(2)} \neq 0$. Substitute finally in the first equation to produce x_1, which is possible if $a_{11} \neq 0$. This completes the back-substitution and we have obtained the required answers.

* U is also referred to as an *upper triangle*.

Example

As an example, let us solve the equations

$$\begin{aligned} x_1 - 3x_2 + 2x_3 &= -12\\ x_1 + 2x_2 + x_3 &= 5\\ -x_1 - 3x_2 - 3x_3 &= -4. \end{aligned}$$

The whole of the arithmetic appears as follows.

$$\begin{array}{lrrrr} & & A & & \mathbf{b}\\ & 1 & -3 & 2 & -12\\ m_{21} = -1 & 1 & 2 & 1 & 5\\ m_{31} = 1 & -1 & -3 & -3 & -4 \end{array}$$

$$\begin{array}{lrrrr} & & A^{(2)} & & \mathbf{b}^{(2)}\\ & 1 & -3 & 2 & -12\\ & 0 & 5 & -1 & 17\\ m_{32} = \frac{6}{5} & 0 & -6 & -1 & -16 \end{array}$$

$$\begin{array}{rrrr} & A^{(3)} = U & & \mathbf{b}^{(3)} = \mathbf{c}\\ 1 & -3 & 2 & -12\\ 0 & 5 & -1 & 17\\ 0 & 0 & -\frac{11}{5} & \frac{22}{5}. \end{array}$$

This completes the elimination.

For the back-substitution, from the last row of $U\mathbf{x} = \mathbf{c}$ we obtain

$$-\tfrac{11}{5}x_3 = \tfrac{22}{5},$$

which gives $x_3 = -2$.

From the second row, $5x_2 - x_3 = 17$, i.e. $5x_2 = 15$, which gives

$$x_2 = 3,$$

and from the first row,

$$x_1 - 3x_2 + 2x_3 = -12,$$

which gives

$$x_1 = 1.$$

Notice a connection between this method and the calculation of the Hermite normal form discussed in *Unit 3*. In the latter process, after the production of $A^{(2)}$ we should not only eliminate the $a_{32}^{(2)}$ term using the multiplier m_{32}, but we should also eliminate the $a_{12}^{(2)}$ term (which is so far just the original a_{12}), by adding a multiple m_{12} of row 2 of $A^{(2)}$ to the first row of $A^{(2)}$. With obvious further operations of this kind the arrays have the following appearance.

$$\begin{array}{rrrrr}
 & & A & & \mathbf{b} \\
 & 1 & -3 & 2 & -12 \\
m_{21} = -1 & 1 & 2 & 1 & 5 \\
m_{31} = 1 & -1 & -3 & -3 & -4 \\
 & & A^{(2)} & & \mathbf{b}^{(2)} \\
m_{12} = \frac{3}{5} & 1 & -3 & 2 & -12 \\
 & 0 & 5 & -1 & 17 \\
m_{32} = \frac{6}{5} & 0 & -6 & -1 & -16 \\
 & & A^{(3)} & & \mathbf{b}^{(3)} \\
m_{13} = \frac{7}{11} & 1 & 0 & \frac{7}{5} & -\frac{9}{5} \\
m_{23} = -\frac{5}{11} & 0 & 5 & -1 & 17 \\
 & 0 & 0 & -\frac{11}{5} & \frac{22}{5} \\
 & & A^{(4)} & & \mathbf{b}^{(4)} \\
 & 1 & 0 & 0 & 1 \\
 & 0 & 5 & 0 & 15 \\
 & 0 & 0 & -\frac{11}{5} & \frac{22}{5}
\end{array}$$

Finally, dividing by the diagonal elements we have the Hermite normal form

$$\begin{array}{rrrr}
 & A^{(5)} & & \mathbf{b}^{(5)} \\
1 & 0 & 0 & 1 \\
0 & 1 & 0 & 3 \\
0 & 0 & 1 & -2
\end{array}$$

and the solution of the linear equations appears in the last column. The reduction of A to Hermite normal form produces essentially a diagonal matrix (and the corresponding elimination method is called the *Jordan method*), for which no back-substitution is needed.

In practice we prefer the Gauss method because it involves fewer numerical operations. For one set of linear equations with a matrix of order n, it is known that the Gauss method requires the computation of n reciprocals, $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ multiplications, and $\frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n$ additions. The Jordan method, on the other hand, requires n reciprocals, $\frac{1}{2}n^3 + n^2 - \frac{1}{2}n$ multiplications, and $\frac{1}{2}n^3 - \frac{1}{2}n$ additions. In calculating the number of operations involved in the two methods, we have assumed that all possible economies are made. For example, in computing the multiplier m_{32}, say, we first evaluate $1/a_{22}^{(2)}$ and retain this number for final use in the back-substitution, or at the stage of producing $\mathbf{b}^{(5)}$ in the Jordan method.

In both methods, of course, there are minor variations. For example, we could divide all the elements of a relevant row of $A^{(k)}$ and $\mathbf{b}^{(k)}$ by the diagonal element before performing the elimination. At the stage $A^{(2)}$, $\mathbf{b}^{(2)}$ of our illustrative example, we had

$$\begin{array}{rrrr}
 & A^{(2)} & & \mathbf{b}^{(2)} \\
1 & -3 & 2 & -12 \\
0 & 5 & -1 & 17 \\
0 & -6 & -1 & -16
\end{array}$$

which we could replace by

$$\begin{array}{rrrcr} & \bar{A}^{(2)} & & & \bar{\mathbf{b}}^{(2)} \\ 1 & -3 & 2 & & -12 \\ 0 & 1 & -\frac{1}{5} & & \frac{17}{5} \\ 0 & -6 & -1 & & -16 \end{array}$$

and perform the elimination with the multipliers given immediately as the other elements in the second column, below the second row in the Gauss method and both above and below in the Jordan method.

This is the technique actually used in our work on the Hermite normal form, but in practice it involves more storage and more arithmetic (and therefore more rounding errors which we shall have to consider in due course). Experience and simple counting of the number of numerical operations reveal that our first illustrated method is much better in these respects.

Exercise

Use the Gauss method to solve the equations $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 4 & -2 & 3 \\ -2 & 4 & 2 \\ 3 & -1 & 2 \end{bmatrix} \quad \text{and } \mathbf{b} = \begin{bmatrix} 2 \\ -4 \\ 3 \end{bmatrix}$$

Solution

Laying out the calculation as in the previously worked example, we get

$$\begin{array}{lrrrcr} & & A & & & \mathbf{b} \\ & 4 & -2 & 3 & & 2 \\ m_{21} = \frac{1}{2} & -2 & 4 & 2 & & -4 \\ m_{31} = -\frac{3}{4} & 3 & -1 & 2 & & 3 \end{array}$$

$$\begin{array}{lrrrcr} & & A^{(2)} & & & \mathbf{b}^{(2)} \\ & 4 & -2 & 3 & & 2 \\ & 0 & 3 & \frac{7}{2} & & -3 \\ m_{32} = -\frac{1}{6} & 0 & \frac{1}{2} & -\frac{1}{4} & & \frac{3}{2} \end{array}$$

$$\begin{array}{rrrcr} & A^{(3)} = U & & & \mathbf{b}^{(3)} = \mathbf{c} \\ 4 & -2 & 3 & & 2 \\ 0 & 3 & \frac{7}{2} & & -3 \\ 0 & 0 & -\frac{5}{6} & & 2 \end{array}$$

We then apply back-substitution to $U\mathbf{x} = \mathbf{c}$:

$$-\tfrac{5}{6}x_3 = 2, \qquad x_3 = -\tfrac{12}{5}$$

$$3x_2 + \tfrac{7}{2}x_3 = -3,\ 3x_2 = \tfrac{27}{5}, \qquad x_2 = \tfrac{9}{5}$$

$$4x_1 - 2x_2 + 3x_3 = 2,\ 4x_1 = \tfrac{64}{5},\ x_1 = \tfrac{16}{5}$$

Matrix Inversion

If A is a non-singular 3×3 matrix, then the matrix X in

$$AX = I$$

is the inverse A^{-1} of A. We can rewrite this equation in the form

$$A[\mathbf{x}^{(1)}\ \mathbf{x}^{(2)}\ \mathbf{x}^{(3)}] = [\mathbf{i}^{(1)}\ \mathbf{i}^{(2)}\ \mathbf{i}^{(3)}]$$

where $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(3)}$ are the columns of X and $\mathbf{i}^{(1)}$, $\mathbf{i}^{(2)}$, $\mathbf{i}^{(3)}$ are the columns of I. Thus, we can find A^{-1} by solving the three sets of three simultaneous equations:

$$A\mathbf{x}^{(r)} = \mathbf{i}^{(r)} \qquad (r = 1, 2, 3)$$

and we can apply the Gauss method above. We do not, however, solve the three sets of equations singly, but we reduce A to upper triangular form as before, performing the same operations on three right-hand sides, which are the columns of the unit matrix, and then carrying out three separate back-substitutions. The full computation for finding the inverse of

$$A = \begin{bmatrix} 1 & -3 & 2 \\ 1 & 2 & 1 \\ -1 & -3 & -3 \end{bmatrix} \text{ is as follows:}$$

Elimination

$$\begin{array}{lrrrrrr}
 & 1 & -3 & 2 & 1 & 0 & 0 \\
m_{21} = -1 & 1 & 2 & 1 & 0 & 1 & 0 \\
m_{31} = 1 & -1 & -3 & -3 & 0 & 0 & 1 \\
 & 1 & -3 & 2 & 1 & 0 & 0 \\
 & 0 & 5 & -1 & -1 & 1 & 0 \\
m_{32} = \frac{6}{5} & 0 & -6 & -1 & 1 & 0 & 1 \\
 & 1 & -3 & 2 & 1 & 0 & 0 \\
 & 0 & 5 & -1 & -1 & 1 & 0 \\
 & 0 & 0 & -\frac{11}{5} & -\frac{1}{5} & \frac{6}{5} & 1 \\
 & & & & (1) & (2) & (3)
\end{array}$$

Back-substitution

Taking columns (1), (2) and (3) in turn yields:

(1) $x_3^{(1)} = \frac{1}{11},\ x_2^{(1)} = -\frac{2}{11},\ x_1^{(1)} = \frac{3}{11}$

(2) $x_3^{(2)} = -\frac{6}{11},\ x_2^{(2)} = \frac{1}{11},\ x_1^{(2)} = \frac{15}{11}$

(3) $x_3^{(3)} = -\frac{5}{11},\ x_2^{(3)} = -\frac{1}{11},\ x_1^{(3)} = \frac{7}{11}$

The inverse has columns $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(3)}$ and is therefore

$$A^{-1} = \frac{1}{11}\begin{bmatrix} 3 & 15 & 7 \\ -2 & 1 & -1 \\ 1 & -6 & -5 \end{bmatrix}$$

8.2.2 The Jordan Method

The Jordan method for inversion, essentially the same as the production of the Hermite normal form, avoids back-substitution. When the matrix on the left has been transformed to the unit matrix, then that on the right is the inverse of the original matrix. The full computation for finding the inverse of

$$A = \begin{bmatrix} 1 & -3 & 2 \\ 1 & 2 & 1 \\ -1 & -3 & -3 \end{bmatrix}$$

is as follows:

$$\begin{array}{lrrrrrrr}
 & 1 & -3 & 2 & \quad & 1 & 0 & 0 \\
m_{21} = -1 & 1 & 2 & 1 & & 0 & 1 & 0 \\
m_{31} = 1 & -1 & -3 & -3 & & 0 & 0 & 1 \\
 & & & & & & & \\
m_{12} = \frac{3}{5} & 1 & -3 & 2 & & 1 & 0 & 0 \\
 & 0 & 5 & -1 & & -1 & 1 & 0 \\
m_{32} = \frac{6}{5} & 0 & -6 & -1 & & 1 & 0 & 1 \\
 & & & & & & & \\
m_{13} = \frac{7}{11} & 0 & 0 & \frac{7}{5} & & \frac{2}{5} & \frac{3}{5} & 0 \\
m_{23} = -\frac{5}{11} & 0 & 5 & -1 & & -1 & 1 & 0 \\
 & 0 & 0 & -\frac{11}{5} & & -\frac{1}{5} & \frac{6}{5} & 1 \\
 & & & & & & & \\
 & 1 & 0 & 0 & & \frac{3}{11} & \frac{15}{11} & \frac{7}{11} \\
 & 0 & 5 & 0 & & -\frac{10}{11} & \frac{5}{11} & -\frac{5}{11} \\
 & 0 & 0 & -\frac{11}{5} & & -\frac{1}{5} & \frac{6}{5} & 1 \\
 & & & & & & & \\
 & 1 & 0 & 0 & & \frac{3}{11} & \frac{15}{11} & \frac{7}{11} \\
 & 0 & 1 & 0 & & -\frac{2}{11} & \frac{1}{11} & -\frac{1}{11} \\
 & 0 & 0 & 1 & & \frac{1}{11} & -\frac{6}{11} & -\frac{5}{11}
\end{array}$$

This is the method already introduced in *Unit 3*.

An interesting fact is that *for inversion* our descriptions of the Gauss and Jordan (Hermite normal form) methods involve identical amounts of arithmetic: n reciprocals, $n^3 - 1$ multiplications and $n^3 - 2n^2 + n$ additions.

8.2.3 Essential Row Interchanges

In the Gauss elimination process we reduced A to U by adding multiples of the first row to all the others to reduce to zero the column 1 elements in rows 2, 3, ..., n. In this process the element in the first row and column is called the *pivot*, and the process is possible only if the pivot is non-zero. After the first reduction to $A^{(2)}$, we continue with the new $a_{22}^{(2)}$ as pivot, to eliminate the new column 2 elements in rows 3, 4, ..., n, and the original pivotal row is unchanged. In other words, we take pivots down the diagonal, reducing A to U, and this is possible if none of the pivots is zero.

The process fails at the first step if a_{11} is zero, and we must then use some other row as pivotal row with its corresponding first element as pivot. It is convenient to interchange the first row of A with the pivotal row, and reduce this "row-permuted" A to the corresponding form $A^{(2)}$. If the new $a_{22}^{(2)}$ is zero, we look for a non-zero column 2 element in rows 3 to n, and interchange the relevant rows. A simple example will illustrate this, concentrating on the reduction of A to an upper triangular matrix U.

$$A$$

$$\begin{array}{cccc} 0 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 2 & 2 & 4 & 0 \\ 1 & 2 & 1 & 1 \end{array}$$

Since $a_{11} = 0$, we must use some other row as first pivotal row. Any one will do, so we interchange the first two, to produce

$$A^{(1)}$$

$$\begin{array}{cccc} 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 2 & 2 & 4 & 0 \\ 1 & 2 & 1 & 1 \end{array}$$

and perform the elimination (only two rows needing attention) to obtain

$$A^{(2)}$$

$$\begin{array}{cccc} 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & -2 \\ 0 & 1 & -1 & 0 \end{array}$$

We can use $a_{22}^{(2)}$ as pivot; there is only the $a_{42}^{(2)}$ term to eliminate, and we obtain

$$A^{(3)}$$

$$\begin{array}{cccc} 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & -2 \\ 0 & 0 & -2 & -1 \end{array}$$

We cannot now use the zero $a_{33}^{(3)}$ as pivot, and we must interchange rows 3 and 4. No further elimination is needed, and we produce the upper triangular form

$$\begin{array}{cccc} 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & -2 & -1 \\ 0 & 0 & 0 & -2 \end{array}$$

Exercises

1. In the last example we could at the start have used the third row as pivotal row, interchanging it with the first. Carry out the relevant

arithmetic to produce an upper triangular form, making subsequent row interchanges as necessary.

2. Using the upper triangular form produced in the text immediately preceding Exercise 1, show that det $A = 4$. Verify this by considering the triangular form obtained in Exercise 1.

3. Suppose that in Exercise 1 we interchange rows 1 and 3 of A, then interchange rows 2 and 3 of the new matrix, and finally rows 3 and 4 of this matrix, *before* performing any elimination. Show that the pivots can now be taken on the diagonal and that the computed U is that quoted in Exercise 1.

Solutions

1. The first array becomes

$$\begin{array}{rcccc} & 2 & 2 & 4 & 0 \\ m_{21} = -\frac{1}{2} & 1 & 1 & 2 & 1 \\ m_{31} = 0 & 0 & 1 & 1 & 1 \\ m_{41} = -\frac{1}{2} & 1 & 2 & 1 & 1 \end{array}$$

and elimination produces

$$\begin{array}{cccc} 2 & 2 & 4 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & -1 & 1 \end{array}$$

We cannot use $a_{22}^{(2)}$ as pivot, but $a_{32}^{(2)}$ is a possibility. Interchanging rows 2 and 3 gives

$$\begin{array}{cccc} 2 & 2 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 1 \end{array}$$

and at this stage there is just one elimination to perform. We find the new matrix

$$\begin{array}{cccc} 2 & 2 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -2 & 0 \end{array}$$

Now $a_{33}^{(3)}$ is not a possible pivot, but $a_{43}^{(3)}$ is possible, and we therefore interchange rows 3 and 4, find no further elimination necessary, and record the final upper triangular form

$$U$$

$$\begin{array}{cccc} 2 & 2 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{array}$$

(Note that after the first elimination it would have been possible to use $a_{42}^{(2)}$ as pivot. This would give a different final upper triangular U.)

2. U from the text has been produced from A by two row interchanges, each of which changes the sign of det A, together with addition of multiples of some rows to others, each of which does not change the sign. Hence

$$\begin{aligned} \det A &= (-1)^2 \det U \\ &= \text{product of diagonal terms of } U. \end{aligned}$$

Similarly, the U of Exercise 1 was produced from A in a way that involved three row interchanges, and det $U = -4$. Thus, both methods give det $A = 4$.

3. The first interchange (of rows 1 and 3 of A) gives

$$\begin{array}{cccc} 2 & 2 & 4 & 0 \\ 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \end{array}$$

The next interchange, of rows 2 and 3 of this matrix, gives

$$\begin{array}{cccc} 2 & 2 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 2 & 1 & 1 \end{array}$$

and the final interchange, of rows 3 and 4 of this matrix, gives

$$\begin{array}{cccc} 2 & 2 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \end{array}$$

The standard elimination gives successive matrices

$$\bar{A}^{(1)}$$

$$\begin{array}{r|cccc} & 2 & 2 & 4 & 0 \\ m_{21} = 0 & 0 & 1 & 1 & 1 \\ m_{31} = -\frac{1}{2} & 1 & 2 & 1 & 1 \\ m_{41} = -\frac{1}{2} & 1 & 1 & 2 & 1 \end{array}$$

$$\bar{A}^{(2)}$$

$$\begin{array}{r|cccc} & 2 & 2 & 4 & 0 \\ & 0 & 1 & 1 & 1 \\ m_{32} = -1 & 0 & 1 & -1 & 1 \\ m_{42} = 0 & 0 & 0 & 0 & 1 \end{array}$$

and the final form is

$$\bar{A}^{(3)} = U$$

$$\begin{array}{cccc} 2 & 2 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{array}$$

Such interchanges as those above were used in the production of the Hermite normal form. We note, here as well as in the work on the Hermite normal form, that if at any stage there is no possible candidate for pivot then the matrix is singular. One might wish to do something else (such as to continue with the production of the Hermite normal form), but so far as the solution of linear equations is concerned we know that we cannot then find a unique solution and that there may not be a solution at all. This is somewhat unlikely in practical work and we shall not deal with it in this unit.

The interchanges above were *essential*, since without them we could not perform the relevant elimination. When we perform an error analysis of the method, to detect the possibility of induced instability, and then look to see if we can produce a more stable method, we shall learn the interesting fact that selected *non-essential* interchanges become very important and are in practice necessary to produce a guaranteed stable method.

We shall consider this in Section 8.3, but for the rest of this section we examine the matrix equivalent of our elimination method, discovering in the process that the same technique can be considered in a completely different way in which the idea of elimination never appears explicitly.

8.2.4 Matrix Equivalent of Gauss Elimination: Matrix Decomposition

We know from previous work (*Unit 3*, for example) that all the operations performed in reducing A to upper triangular form U can be expressed in terms of matrix operations on A. Suppose, for example, that we *can* take pivots down the diagonal. Then it can easily be seen, using a 3×3 matrix for illustration as usual, that we first form

$$A^{(2)} = J_1 A = \overset{J_1}{\begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{bmatrix}} \overset{A}{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}}$$

where $m_{21} = -a_{21}/a_{11}$, $m_{31} = -a_{31}/a_{11}$. The premultiplying matrix J_1 is lower triangular with unit elements on the diagonal. Such a matrix is *unit lower triangular.*

At the next and final stage in our example, we have

$$A^{(3)} = U = J_2 J_1 A, \quad J_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_{32} & 1 \end{bmatrix}$$

where $m_{32} = -a_{32}^{(2)}/a_{22}^{(2)}$.

This extends in an obvious way to the general case: the result is

$$J_{n-1} J_{n-2} \quad \cdots \quad J_2 J_1 A = U,$$

where U is upper triangular and each J_r is unit lower triangular.

The product of (unit) lower triangular matrices is (unit) lower triangular, and the inverse of a (unit) lower triangular matrix is also (unit) lower triangular (see the Exercise below), and it follows that we can regard the elimination process as the equivalent of the *triangular decomposition* of a matrix in the form

$$A = LU,$$

where L is a unit lower triangle and U an upper triangle.

It is of interest to determine the nature of L in terms of the J_r. We have

$$L = (J_{n-1} \, J_{n-2} \cdots J_2 \, J_1)^{-1} = J_1^{-1} \, J_2^{-1} \cdots J_{n-1}^{-1}.$$

For the case of order three we easily verify that

$$J_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ {-}{-}m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{bmatrix},$$

which is identical with J_1 except for sign changes in the off-diagonal elements. Similarly

$$J_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{bmatrix}.$$

Finally, we find that the product $L = J_1^{-1} \, J_2^{-1}$ has the interesting form

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & -m_{32} & 1 \end{bmatrix},$$

none of the m_{rs} being multiplied together!

It follows that, in the decomposition $A = LU$, the U matrix is the form obtained from the elimination process, and the L matrix is formed from the negatives of the multipliers in the elimination process.

Exercise

Verify for 3 × 3 matrices the statements made above:

(i) the product of (unit) lower triangular matrices is (unit) lower triangular;

(ii) the inverse of a (unit) lower triangular matrix is also (unit) lower triangular.

Solution

(i)
$$\begin{bmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ l & 1 & 0 \\ m & n & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ a+l & 1 & 0 \\ b+cl+m & c+n & 1 \end{bmatrix}$$

(ii) Employing the method of *Unit 3*, with the notation of this unit:

$$\begin{array}{lccccccc}
 & 1 & 0 & 0 & & 1 & 0 & 0 \\
m_{21} = -a & a & 1 & 0 & & 0 & 1 & 0 \\
m_{31} = -b & b & c & 1 & & 0 & 0 & 1 \\
 & 1 & 0 & 0 & & 1 & 0 & 0 \\
 & 0 & 1 & 0 & & -a & 1 & 0 \\
m_{32} = -c & 0 & c & 1 & & -b & 0 & 1 \\
 & 1 & 0 & 0 & & 1 & 0 & 0 \\
 & 0 & 1 & 0 & & -a & 1 & 0 \\
 & 0 & 0 & 1 & & ac-b & -c & 1
\end{array}$$

8.2.5 Compact Elimination

The discussion in the previous sub-section gives us a method for solving linear equations which in explicit terms has nothing to do with elimination. Consider, for example, a 3 × 3 case and write

$$\underset{L}{\begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}}\underset{U}{\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}} = \underset{A}{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}}$$

By equating the product of the rth row, $L(r)$, of L and sth column, $U(s)$, of U with the element a_{rs} we can calculate successively the elements of L and U. If $L(r)U(s)$ denotes the product of the rth row of L and the sth column of U, we have

$$L(1)U(1) = u_{11} = a_{11}$$
$$L(1)U(2) = u_{12} = a_{12}$$
$$L(1)U(3) = u_{13} = a_{13}$$
$$L(2)U(1) = l_{21}u_{11} = a_{21}$$

whence

$$l_{21} = a_{21}/u_{11},$$

if $u_{11} = a_{11} \neq 0$.

$$L(2)U(2) = l_{21}u_{12} + u_{22} = a_{22}$$

whence

$$u_{22} = a_{22} - l_{21} \times u_{12}$$

which gives u_{22}, and so on.

The computations are performed to produce, in order, the first row of U, second row of L, second row of U, and so on.

Exercise

Find the triangular decomposition of

$$A = \begin{bmatrix} 1 & -3 & 2 \\ 1 & 2 & 1 \\ -1 & -3 & -3 \end{bmatrix}$$

and compare the results with those obtained for A by the Gauss elimination method in sub-section 8.2.1.

Solution

Let

$$\begin{bmatrix} 1 & -3 & 2 \\ 1 & 2 & 1 \\ -1 & -3 & -3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

Then ordinary matrix multiplication gives us $1 = u_{11}$, $-3 = u_{12}$, $2 = u_{13}$

$$\begin{aligned} 1 &= l_{21}u_{11} && \text{whence} \quad l_{21} = 1 \\ 2 &= l_{21}u_{12} + u_{22} && \text{whence} \quad u_{22} = 5 \\ 1 &= l_{21}u_{13} + u_{23} && \text{whence} \quad u_{23} = -1 \\ -1 &= l_{31}u_{11} && \text{whence} \quad l_{31} = -1 \\ -3 &= l_{31}u_{12} + l_{32}u_{22} && \text{whence} \quad l_{32} = -\frac{6}{5} \\ -3 &= l_{31}u_{13} + l_{32}u_{23} + u_{33} && \text{whence} \quad u_{33} = -\frac{11}{5}. \end{aligned}$$

We have, therefore

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -\frac{6}{5} & 1 \end{bmatrix} \begin{bmatrix} 1 & -3 & 2 \\ 0 & 5 & -1 \\ 0 & 0 & -\frac{11}{5} \end{bmatrix}$$

All the figures can be found in our previous work (in sub-section 8.2.1) on the matrix A, using Gauss elimination.

We can now see quite easily when the method illustrated above will fail. It is clear that the product of successive leading submatrices * of L and U is equal to the successive leading submatrices of A, that is

$$[1][u_{11}] = [a_{11}]$$

$$\begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

etc.

It follows that we cannot compute l_{21} (and hence u_{22}) unless $u_{11} \neq 0$. If $u_{11} \neq 0$, we can produce the second decomposition given above, but can go no further if $u_{22} = 0$.

Now $u_{11} = a_{11} = \det [a_{11}]$

$$u_{11}\, u_{22} = \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

and we have therefore shown that the method will succeed only if successive leading submatrices of A are non-singular. We can complete the decomposition if the first $n-1$ such matrices are non-singular, but if A is singular then the final element u_{nn} will be zero.

* The meaning of the term *leading submatrix* should be clear from what follows.

In fact the essential interchanges of sub-section 8.2.3 are designed so that some row-permuted form of A has non-singular leading sub-matrices, and it follows that if A itself is non-singular we can always form the triangular decomposition of this row-permuted A. If A is singular with rank $n-1$ we can also perform the decomposition, though the last element of U will be zero.

It is not quite obvious how this "interchanging" process can be incorporated into the automatic computation of L and U, and we shall not go into details about this. The possibility, however, is revealed in the following exercise.

Exercise

Consider the matrix $A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 2 & 2 & 4 & 0 \\ 1 & 2 & 1 & 1 \end{bmatrix}$

In sub-section 8.2.3 we carried out on A an elimination process with interchanges, first interchanging rows 1 and 2 and subsequently rows 3 and 4. Write down the matrix derived from A by making these interchanges in advance, compute the triangular decomposition of this matrix, and compare the results with those of the elimination method.

Solution

The row-permuted matrix and the triangular decomposition are

$$\begin{bmatrix} 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 2 & 2 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & -2 & -1 \\ 0 & 0 & 0 & -2 \end{bmatrix}$$

and no $u_{rr} = 0$. Observe that U is the same as the upper triangle of the elimination process with row interchanges.

To complete the solution of linear equations, using the idea of triangular decomposition, we can write

$$[A, \mathbf{b}] = L[U, \mathbf{c}],$$

the vector $\mathbf{c}$ being obtained from $\mathbf{b}$ by the same process by which the columns of U are obtained from A. Alternatively, and equivalently, we can compute $\mathbf{c}$, having found L, by *forward* substitution in the equations $L\mathbf{c} = \mathbf{b}$, written explicitly as

$$\begin{aligned} c_1 \qquad\qquad\quad &= b_1 \\ l_{21}c_1 + \quad c_2 \qquad &= b_2 \\ l_{31}c_1 + l_{32}c_2 + c_3 &= b_3, \quad \text{etc.} \end{aligned}$$

The vector $\mathbf{c}$ is of course the right-hand vector which is produced from $\mathbf{b}$ in the standard Gauss elimination process. The final step in the solution obtains $\mathbf{x}$ from $U\mathbf{x} = \mathbf{c}$ by back-substitution, as in the Gauss elimination method.

Exercise

Solve the equations

$$\begin{aligned} x_1 - 3x_2 + 2x_3 &= -12 \\ x_1 + 2x_2 + x_3 &= \quad 5 \\ -x_1 - 3x_2 - 3x_3 &= -\ 4 \end{aligned}$$

by computing L, U, and $\mathbf{c}$ from the equations $[A, \mathbf{b}] = L[U, \mathbf{c}]$, and obtain $\mathbf{x}$ by back-substitution in $U\mathbf{x} = \mathbf{c}$. Compare the results, at every stage in the computation, with those of the first example of sub-section 8.2.1.

Solution

We find

$$\overset{[A,\,\mathbf{b}]}{\begin{bmatrix} 1 & -3 & 2 & -12 \\ 1 & 2 & 1 & 5 \\ -1 & -3 & -3 & -4 \end{bmatrix}}$$

$$= \overset{[L]}{\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -\dfrac{6}{5} & 1 \end{bmatrix}} \overset{[U,\,\mathbf{c}]}{\begin{bmatrix} 1 & -3 & 2 & -12 \\ 0 & 5 & -1 & 17 \\ 0 & 0 & -\dfrac{11}{5} & \dfrac{22}{5} \end{bmatrix}}$$

and the back-substitution follows as before.

A noticeable feature of this *compact elimination method* is that we record only the elements of L, which are the negatives of the multipliers in the Gauss elimination process, the elements of U, which is the final upper triangular matrix $U = A^{(3)}$ of the Gauss elimination process, and the final right-hand vector $\mathbf{c} = \mathbf{b}^{(3)}$ of the Gauss elimination method. In other words, we apparently compute and certainly record only the *pivotal rows* of the Gauss elimination method. What, you might well ask, has happened to a non-pivotal computed row such as the third row of the matrix $A^{(2)}$ and vector $\mathbf{b}^{(2)}$? An answer to this is indicated in the following exercise.

Exercise

In the previous exercise, the last element of the vector $\mathbf{c}$, which is the element $\mathbf{b}_3^{(3)}$ of the Gauss elimination method, is obtained from the equation

$$l_{31}\, c_1 + l_{32}\, c_2 + c_3 = b_3,$$

that is

$$c_3 = b_3 - l_{31}\, c_1 - l_{32}\, c_2 .$$

Verify that

$$b_3 - l_{31}\, c_1 = b_3^{(2)}$$

(of the Gauss elimination process).

Solution

$$b_3 - l_{31}\, c_1 = -4 - (-1)(-12) = -16 = b_3^{(2)}.$$

In fact, all the unrecorded numbers of the Gauss elimination method are the partial sums of the formula for the final element in the relevant place in the array. The practical importance of this result will become apparent when we perform an error analysis of the Gauss elimination method. This we proceed to do, demonstrating in the process that the Gauss elimination method as we have described it, with interchanges only where *essential*, can, in practice, suffer from acute induced instability when we use inexact computer arithmetic.

8.2.6 Summary of Section 8.2

In this section we defined the terms

Gauss elimination method	(page C 12)	* * *
upper triangular matrix	(page C 12)	* * *
back-substitution	(page C 12)	* * *
Jordan method	(page C 17)	* * *
pivot	(page C 18)	* * *
essential row interchanges	(page C 20)	* *
unit lower triangular matrix	(page C 21)	* *
triangular decomposition of a matrix	(page C 21)	* *
compact elimination method	(page C 25)	*

We introduced the notation

m_{ij}	(page C 12)
$\mathbf{b}^{(k)}$	(page C 12)
$A^{(k)}$	(page C 12)
U	(page C 12)
L	(page C 21)
$L(r)$	(page C 22)
$U(s)$	(page C 22)

Techniques

1. Solve $A\mathbf{x} = \mathbf{b}$ and invert A by the Gauss elimination method and by the Jordan method. *
2. Determine the triangular decomposition of A. * * *
3. Solve $A\mathbf{x} = \mathbf{b}$ by the compact elimination method. * * *
4. Evaluate det A by reducing A to triangular form. * * *

8.3 INDUCED INSTABILITY OF GAUSS ELIMINATION

8.3.1 A Numerical Example

Following our standard pattern, the next thing we should ask is whether or not the Gauss elimination method suffers from induced instability when, as is almost invariably the case, we have to use a computer on a practical problem.

First we give an example to show that this is undoubtedly possible, and we use for this purpose a hypothetical four-decimal-digit floating-point computer. The equations are

$$\begin{aligned}
-1.414x_1 + x_2 \qquad\qquad\qquad\qquad &= 0.1000\\
x_1 - 1.414x_2 + x_3 \qquad\qquad &= 0.1000\\
x_2 - 1.414x_3 + x_4 &= 0.1000\\
x_3 - 1.414x_4 &= 0.1000
\end{aligned}$$

which may have come from a boundary-value problem involving a recurrence relation of second order, introduced in *Unit 7, Recurrence Relations.*

At each stage there is only one number to eliminate, and we can write down the upper triangular form in the process of computing it. The multipliers are written on the left, and we find

$$\begin{array}{lrcl}
0.7072 & -1.414x_1 + x_2 \qquad\qquad\qquad & = & 0.1000\\
1.415 & -0.7068x_2 + x_3 \qquad\quad & = & 0.1707\\
-1000 & 0.0010x_3 + x_4 & = & 0.3415\\
 & -1001x_4 & = & -341.4
\end{array}$$

Back-substitution gives

$$x_4 = 0.3411,\ x_3 = 0.4000,\ x_2 = 0.3244,\ x_1 = 0.1587.$$

A glance at the original equations reveals a symmetry $x_1 = x_4$, $x_2 = x_3$, and our answers are clearly far from the truth. The problem, incidentally, is not particularly ill-conditioned, and in fact the method has exhibited considerable induced instability: somewhat surprising, since this is really a very small problem!

Let us do some non-essential interchanges, not *theoretically* necessary, and re-order the equations to read

$$\begin{aligned}
-1.414x_1 + x_2 \qquad\qquad\qquad\qquad &= 0.1000\\
x_2 - 1.414x_3 + x_4 &= 0.1000\\
x_3 - 1.414x_4 &= 0.1000\\
x_1 - 1.414x_2 + x_3 \qquad\qquad &= 0.1000
\end{aligned}$$

It is perhaps now more convenient to record separately the successive "reduced" sets of equations. We have (with pivots emboldened)

Multipliers	*Matrices*				*Vectors*
	−1.414	1	0	0	0.1000
	0	1	−1.414	1	0.1000
	0	0	1	−1.414	0.1000
0.7072	1	−1.414	1	0	0.1000
	−1.1414	1	0	0	0.1000
	0	**1**	−1.414	1	0.1000
	0	0	1	−1.414	0.1000
0.7068	0	−0.7068	1	0	0.1707
	−1.414	1	0	0	0.1000
	0	1	−1.414	1	0.1000
	0	0	**1**	−1.414	0.1000
−0.0006	0	0	0.0006	0.7068	0.2414
	−1.414	1	0	0	0.1000
	0	1	−1.414	1	0.1000
	0	0	1	−1.414	0.1000
	0	0	0	0.7076	0.2413

Back-substitution gives

$$x_4 = 0.3410,\ x_3 = 0.5822,\ x_2 = 0.5822,\ x_1 = 0.3410.$$

Symmetry reveals that the answers are almost certainly correct to four decimal figures, and this can be checked by more accurate computation. We have eliminated the induced instability!

8.3.2 Error Analysis of the Gauss Elimination Method (backward error analysis)

In *Unit 7, Recurrence Relations*, we investigated the instability evidenced by computations with recurrence relations. For example, for the problem defined by

$$y_{r+1} = a_r y_r + b_r, \qquad y_0 = \alpha,$$

we observed that a *local* error is made in the floating-point arithmetic computation of any y_{r+1} from a previously computed y_r, and we analysed the effect of those errors to deduce the total error in the computed y_n, say. This total error was compared with the value of y_n which would be produced by exact arithmetic. (We assumed, for this purpose, that all the *data* are known exactly and can be stored exactly.)

This method of analysis, which gives an estimate of the error in the computed result, is called *forward error analysis.* It turns out that this is an extremely complicated operation when applied to the solution of linear equations. We prefer here the easier method of *backward error analysis*, whose aim is not to find the errors in the computed solutions but to discover what *changes in the data* would produce these solutions with *exact* arithmetic. We try to show that our method produces a solution vector $\bar{\mathbf{x}}$ which, while not satisfying exactly the given equations $A\mathbf{x} = \mathbf{b}$, will satisfy exactly the "perturbed" equations $(A + \delta A)\bar{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$. If the elements in the perturbations δA* and $\delta\mathbf{b}$ are small we have *little induced instability*, and any dissatisfaction with the solution can spring only from the *possibility of inherent instability*. By looking at the sizes of perturbations for different methods we evaluate our techniques, retaining those for which

* Note that δA is not "δ times A".

δA and $\delta \mathbf{b}$ have small elements relative to those of A and $\mathbf{b}$ respectively, and discarding those for which these quantities are, or can be, relatively large.

The basic numerical operations in the Gauss elimination process are
(i) the computation of a multiplier, such as $m_{21} = -a_{21}/a_{11}$, and
(ii) the computation of a new element, such as $a_{22}^{(2)} = a_{22} + m_{21}a_{12}$.
In the computation (i) the machine makes a rounding error and stores

$$\bar{m}_{21} = m_{21} + e_{21} = -\frac{a_{21}}{a_{11}} + e_{21} = -\frac{a_{21} - e_{21}a_{11}}{a_{11}}$$

Hence the calculated multiplier would be the *exact* value, using *exact* arithmetic, if a_{21} were replaced by $a_{21} - e_{21}a_{11}$.

In the computation (ii) the machine tries to compute

$$a_{22}^{(2)} = a_{22} + \bar{m}_{21}a_{12},$$

but only succeeds in computing and storing

$$\bar{a}_{22}^{(2)} = a_{22} + \bar{m}_{21}a_{12} + e_{22} = (a_{22} + e_{22}) + \bar{m}_{21}a_{12}.$$

It follows that the stored element would be the correct value, using exact arithmetic, if a_{22} were replaced by $a_{22} + e_{22}$.

We shall later want to find upper bounds for these "perturbations" e_{rs}, but to see how it goes let us do an elimination using one-figure floating-point arithmetic, and compute the perturbed matrix which would give rise exactly to the computed and stored multipliers and upper triangular matrix.

We take

$$A = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ -0.5 & 0.5 & -0.2 \\ 0.7 & 0.3 & 0.9 \end{bmatrix}$$

With $a_{11} = 0.8$ as pivot, the multipliers are $m_{21} = \dfrac{0.5}{0.8}$ and $m_{31} = \dfrac{-0.7}{0.8}$. In our floating-point arithmetic, however, the machine stores $\bar{m}_{21} = 0.6$ and $\bar{m}_{31} = -0.9$. These would be exact if we changed the a_{21} term from -0.5 to -0.48, that is added 0.02 to it, and changed the a_{31} term to 0.72, which is again an addition of 0.02.

Now consider the computation of the new $a_{22}^{(2)}$. We have

$$a_{22}^{(2)} = 0.5 + 0.6 \times 0.2 = 0.62 = 0.6 + 0.02.$$

We store 0.6, and this would be exact if we had changed the original a_{22} from 0.5 to 0.48.

The stored first step of the elimination giving $A^{(2)}$, and the changes $\delta_1 A$ in A which would produce everything exactly, are

$A^{(2)}$			$\delta_1 A$		
0.8	0.2	0.1	0.00	0.00	0.00
0	0.6	−0.1	0.02	−0.02	0.04
0	0.1	0.8	0.02	−0.02	−0.01

We can now ignore the first row and column, and repeat the procedure with the remaining rows and columns.

Starting with

$$\begin{matrix} 0.6 & -0.1 \\ 0.1 & 0.8 \end{matrix}$$

the stored multiplier is -0.2, which would be obtained exactly if the $a_{32}^{(2)} = 0.1$ were changed to 0.12.

The new

$$a_{33}^{(3)} = 0.8 - 0.2(-0.1) = 0.82 = 0.8 + 0.02$$

and this would be produced exactly if we made a change of -0.02 in $a_{33}^{(2)}$.

Writing everything in full, we have the arrays:

multipliers		A				
	0.8	0.2	0.1			
0.6	−0.5	0.5	−0.2			
−0.9	0.7	0.3	0.9			
		$A^{(2)}$			$\delta_1 A$	
	0.8	0.2	0.1	0.00	0.00	0.00
	0	**0.6**	−0.1	0.02	−0.02	0.04
−0.2	0	0.1	0.8	0.02	−0.02	−0.01
		$A^{(3)} = U$			$\delta_2 A$	
	0.8	0.2	0.1	0.00	0.00	0.00
	0	0.6	−0.1	0.00	0.00	0.00
	0	0	0.8	0.00	0.02	−0.02

The important fact is that these perturbations are additive; elimination with exact arithmetic would produce exactly all the multipliers and the final upper triangular form if we started with the matrix $A + \delta_1 A + \delta_2 A$.

Exercise

The easiest way to check this last statement is to express it in terms of the triangular decomposition. Show that the stored L and U satisfy exactly the equation

$$LU = A + \delta_1 A + \delta_2 A.$$

Solution

The matrix $A + \delta_1 A + \delta_2 A$ is

$$\begin{bmatrix} 0.8 & 0.2 & 0.1 \\ -0.5 & 0.5 & -0.2 \\ 0.7 & 0.3 & 0.9 \end{bmatrix} + \begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.02 & -0.02 & 0.04 \\ 0.02 & -0.02 & -0.01 \end{bmatrix}$$

$$+ \begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 \\ 0.00 & 0.02 & -0.02 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ -0.48 & 0.48 & -0.16 \\ 0.72 & 0.30 & 0.87 \end{bmatrix}$$

and it is easy to check that this is *exactly* the product LU of the stored triangular matrices

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -0.6 & 1 & 0 \\ 0.9 & 0.2 & 1 \end{bmatrix}, \; U = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0 & 0.6 & -0.1 \\ 0 & 0 & 0.8 \end{bmatrix}$$

In general, if η is the largest possible absolute value of any element of the matrices $\delta_1 A$ and $\delta_2 A$ and in the general case of the matrices $\delta_1 A$, $\delta_2 A$, $\ldots$, $\delta_{n-1}A$, then the pattern of perturbation in A, given by $\delta_1 A + \delta_2 A + \cdots + \delta_{n-1}A$, is bounded by

$$\eta\begin{bmatrix} 0 & 0 & 0 & \cdots \\ 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ - & - & - & \end{bmatrix} + \eta\begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 1 & 1 & \cdots \\ - & - & - & - & \end{bmatrix}$$

$$+\,\eta\begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 1 & \cdots \\ 0 & 0 & 1 & 1 & \cdots \\ - & - & - & - & \end{bmatrix} + \cdots = \eta\begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 1 & 1 & \cdots \\ 1 & 2 & 2 & 2 & \cdots \\ 1 & 2 & 3 & 3 & \cdots \\ - & - & - & - & \end{bmatrix}$$

the final element being $\eta(n-1)$.

The important point, of course, is the size of η, and this we evaluate with our rules for floating-point arithmetic.

In the computation of the multiplier, $m_{21} = -a_{21}/a_{11}$, for example, we have

$$-\bar{m}_{21} = \mathrm{fl}\,(a_{21}/a_{11}) = \frac{a_{21}}{a_{11}}(1+e),$$

where $e \leqslant 2^{-t}$ in a binary machine of word-length t binary digits. The perturbation in a_{21} is then at most $2^{-t}|a_{11}|$ in absolute value.

Where the formation of a new element is concerned, for example

$$a_{22}^{(2)} = a_{22} + \bar{m}_{21}a_{12},$$

we have

$$\begin{aligned}\bar{a}_{22}^{(2)} &= \mathrm{fl}\,(a_{22} + \bar{m}_{21}a_{12}) \\ &= \{a_{22} + \bar{m}_{21}a_{12}(1+e_1)\}\,(1+e_2),\end{aligned}$$

the e_1 coming from a floating-point multiplication and the e_2 from a floating-point addition, and $|e_1|$, $|e_2| \leqslant 2^{-t}$.

The perturbation $\bar{a}_{22}^{(2)} - (a_{22} + \bar{m}_{21}a_{12})$ in a_{22} is then

$$\bar{a}_{22}^{(2)} - \frac{\bar{a}_{22}^{(2)}}{1+e_2} - \bar{m}_{21}a_{12}e_1 = \bar{a}_{22}^{(2)}e_2 - \bar{m}_{21}a_{12}e_1,$$

where we neglect terms like $(e_2)^2$ which are at most 2^{-2t}. The perturbation is then at most

$$2^{-t}(|\bar{a}_{22}^{(2)}| + |\bar{m}_{21}a_{12}|)$$

in absolute value.

These results give upper bounds for the elements of $\delta_1 A$. For $\delta_2 A$ we have similar results, except that now the a_{rs} in these expressions is replaced by $\bar{a}_{rs}^{(2)}$, and so on.

Our η, the largest of these elements therefore depends on (i) the sizes of the multipliers m_{rs}, and (ii) the sizes of the elements $\bar{a}_{rs}^{(k)}$. The perturbations relative to the elements of the original A, which are obviously the important quantities, depend on the multipliers and the possible relative growths in the elements of successive "reduced" matrices $A^{(k)}$.

The failure of the first method for our example of sub-section 8.3.1 is now quite apparent. There was one very large multiplier of -1000, and this also caused (as it always will) a large growth in one $\bar{a}_{rs}$, the last element of the U matrix being -1001. In the second method the multipliers never exceeded unity, the $\bar{a}_{rs}^{(k)}$ did not grow in size as k increased, and the solution was very good.

It is clear, incidentally, that the corresponding reduction of the right-hand vector $\mathbf{b}$ to its final form $\mathbf{c}$ induces errors of exactly the same kind as those relevant to the last column of A. We have therefore shown that the *elimination* process for $A\mathbf{x} = \mathbf{b}$ reduces the problem exactly to the solution of $U\bar{\mathbf{x}} = \mathbf{c}$, and that $\bar{\mathbf{x}}$ is the exact solution of the perturbed problem

$$(A + \delta A)\bar{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}.$$

The perturbations δA and $\delta\mathbf{b}$ can be arbitrarily large if any multipliers or any of the successive elements in $A^{(k)}$ or $\mathbf{b}^{(k)}$ are relatively large compared with those of A and $\mathbf{b}$.

Exercise

In the first method of sub-section 8.3.1, two stages of the elimination produced the array

$$\begin{array}{rrrrr}
 & A^{(3)} & & & \mathbf{b}^{(3)} \\
-1.414 & 1 & 0 & 0 & 0.1000 \\
0 & -0.7068 & 1 & 0 & 0.1707 \\
0 & 0 & 0.0010 & 1 & 0.3415 \\
0 & 0 & 1 & -1.414 & 0.1000
\end{array}$$

Show that the last stage of the elimination, which produces $A^{(4)}$, alone induces in the coefficient $a_{44} = -1.414$, a perturbation of amount equal to nearly one-third of the value of a_{44}, representing a large degree of induced instability.

Are there any other contributions, in this example, to the perturbation of a_{44}?

Solution

The multiplier at the next stage of the elimination is exactly -1000, and the new value computed in the (4, 4) position is

$$\mathrm{fl}\,(-1000(1) - 1.414) = -1001.$$

The true value of this quantity is -1001.414, and the error, the current contribution to the perturbation induced in a_{44}, is -0.414, nearly one-third of $a_{44} = -1.414$.

There are no other contributions to the perturbation in a_{44}, since $a_{44}^{(k)}$ is the same as a_{44} for $k = 2$ and 3.

8.3.3 A Stable Form of Gauss Elimination

In the elimination process using exact arithmetic, or in the corresponding calculation of the Hermite normal form, we had occasionally to interchange rows to produce a non-zero pivot (essential interchanges). This device is used deliberately in Gauss elimination to ensure stability.

Clearly we can make all the multipliers $m_{rs} \leqslant 1$ in absolute value by choosing as pivot the largest in absolute value of the possible candidates in the relevant column, and interchanging the rows to bring this element into the pivotal position. This also has a beneficial effect on the possible growth of the $\bar{a}_{rs}^{(j)}$. We cannot guarantee that they will not grow, but the growth is limited and in practice tends to be tolerable.

With this method (Gauss elimination with interchanges, sometimes called *pivoting*), we can now give an absolute error bound for the induced perturbation.

If we look at the induced perturbations, take $|\bar{m}_{rs}| \leqslant 1$, and let k be the largest in absolute value of all the elements in all the reduced matrices (including the original A), then, using the results established in sub-section 8.3.2, we can now see that the pattern of perturbation is certainly bounded by

$$2 \times 2^{-t} \times k \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 1 & 1 & \cdots \\ 1 & 2 & 2 & 2 & \cdots \\ 1 & 2 & 3 & 3 & \cdots \end{bmatrix}$$

That is, we have produced a U which would be obtained by exact arithmetic from the perturbed matrix $A + \delta A$, where the elements of δA are bounded as shown, and for which the largest coefficient is no greater than

$$(\delta A)_{nn} \leqslant 2(n-1)2^{-t}k.$$

Completing the story, by considering the effect on the right-hand vector $\mathbf{b}$, the elimination has produced the upper triangular equations $U\bar{\mathbf{x}} = \mathbf{c}$, and this would have been produced exactly from $(A + \delta A)\bar{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$, where the elements of maximum size in the perturbations are

$$(\delta A)_{nn} \leqslant 2(n-1)2^{-t}k, \; (\delta \mathbf{b})_n \leqslant 2(n-1)2^{-t}l,$$

where l is the largest in absolute value of all the elements on the right-hand side of the original or any "reduced" equations.

Since the elimination operation involves roughly $\frac{1}{3}n^3$ additions and multiplications, this is an extremely satisfactory result. If k is not significantly larger than the original elements of A, then even for $n = 10^3$ the perturbation affects at most the last four decimal digits of our stored numbers. Quite commonly our binary machine stores the equivalent of ten or eleven decimal digits, so that these induced perturbations are likely to be far smaller, even for this large problem, than any inherent uncertainties in the data. In fact they will generally be very much smaller, since our bounds considerably over-estimate the precise perturbations.

Exercises

1. Consider the solution of linear equations in which the matrix is

$$A = \begin{bmatrix} -\sqrt{2} & 1 & 0 & 0 \\ 1 & -\sqrt{2} & 1 & 0 \\ 0 & 1 & -\sqrt{2} & 1 \\ 0 & 0 & 1 & -\sqrt{2} \end{bmatrix}.$$

By considering the values of the determinants of the first three principal submatrices of A, using exact arithmetic, show that when computer arithmetic is used some pivoting would be essential to avoid induced instability.
(*Hint* Relate the determinants of the triangular decomposition of A, as discussed in sub-section 8.2.5.)
2. By a similar process show that if the rows of A are permuted to the order 1, 3, 4, 2, then pivoting is likely to be unnecessary.

Solutions

1. By relating the elimination to the triangular decomposition of A, we know that the diagonal elements of the U matrix satisfy the equations

$$u_{11} = \det[a_{11}] = -\sqrt{2}$$

$$u_{11}\,u_{22} = \det\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \det\begin{bmatrix} -\sqrt{2} & 1 \\ 1 & -\sqrt{2} \end{bmatrix}$$

$$= 2 - 1 = 1$$

$$u_{11}\,u_{22}\,u_{33} = \det\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$= \det\begin{bmatrix} -\sqrt{2} & 1 & 0 \\ 1 & -\sqrt{2} & 1 \\ 0 & 1 & -\sqrt{2} \end{bmatrix}$$

$$= -\sqrt{2}(2-1) + 1(\sqrt{2}) = 0.$$

With computer arithmetic this computed value is unlikely to be zero but would be very small. The computed diagonal element u_{33} is therefore very small, producing a large next multiplier and hence induced instability.

2. In the new ordering we have

$$\bar{A} = \begin{bmatrix} -\sqrt{2} & 1 & 0 & 0 \\ 0 & 1 & -\sqrt{2} & 1 \\ 0 & 0 & 1 & -\sqrt{2} \\ 1 & -\sqrt{2} & 1 & 0 \end{bmatrix}.$$

The determinants of the first three leading submatrices are

$$-\sqrt{2},\ -\sqrt{2},\ -\sqrt{2}$$

(the product of the diagonal terms, since all these matrices are upper triangular). None of these is small, so that there is no large multiplier and induced instability is unlikely.

These results explain the phenomena of the example of sub-section 8.3.1.

We should, of course, finally investigate whether there is any further instability in the solution by back-substitution of the equations $U\bar{\mathbf{x}} = \mathbf{c}$. We merely state that we can use a similar process of backward error analysis to show that the finally computed solution $\bar{\bar{\mathbf{x}}}$ is the exact solution of another perturbed form

$$(A + \overline{\delta A})\bar{\bar{\mathbf{x}}} = \mathbf{b} + \overline{\delta\mathbf{b}},$$

where, after pivoting in the elimination, the bounds on $\overline{\delta A}$ and $\overline{\delta\mathbf{b}}$ are not more than twice those for the perturbations induced solely by the elimination. The stability of the complete Gauss elimination method is then assured.

It is worth observing that if we used the $\bar{A} = LU$ decomposition, where $\bar{A}$ is the row permutation of A which guarantees that all the multipliers and therefore all the elements of L are numerically less than 1, then with just a little more accurate arithmetic we can reduce the perturbations even further, to spectacularly small values!

We observed at the end of sub-section 8.2.5 that the $A = LU$ method produces u_{nn}, for example, from the formula

$$a_{nn}^{(n)} = u_{nn} = a_{nn} - l_{n1}\, u_{1n} - l_{n2}\, u_{2n} \cdots - l_{n,n-1}\, u_{n-1,n}$$

and that the partial sums

$$a_{nn} - l_{n1}\, u_{1n}$$
$$(a_{nn} - l_{n1}\, u_{1n}) - l_{n2}\, u_{2n}$$
$$\{(a_{nn} - l_{n1}\, u_{1n}) - l_{n2}\, u_{2n}\} - l_{n3}\, u_{3n}$$

etc.

are the elements $a_{nn}^{(2)}$, $a_{nn}^{(3)}$, etc. in the successive "reduced" matrices. Now, we have seen on page 33 that the perturbation induced in a_{nn} by the elimination is something like $2 \times 2^{-t} \times k \times (n-1)$, and the factor $n - 1$ comes from the fact that we actually evaluate u_{nn} by forming each partial sum and rounding it before adding the next term. There are $n - 1$ such rounding errors.

If it were possible, as it is in some machines, to evaluate u_{nn} by doing the multiplications and additions accurately (in "double-length" arithmetic) and doing just one final rounding to single length (so that we are doing a sort of partial double-length arithmetic), then we remove the factor $n - 1$ and all similar factors in the perturbation array, and reduce the maximum perturbation to $2 \times 2^{-t} \times k$, which is very small and *independent* of n. This method, in fact, is used in the best computer programs when high accuracy is needed in the solution.

But this raises another question. We may have a stable method, but can we say how good our computed solution is, and can we from our technique get any idea about the degree of ill-conditioning of the problem? This we discuss in the next section.

8.3.4 Summary of Section 8.3

In this section we defined the terms

forward error analysis	(page C 28)	* *
backward error analysis	(page C 28)	* *
pivoting	(page C 32)	* *

We introduced the notation

δA	(page C 28)
$\delta \mathbf{b}$	(page C 28)
$\overline{m}_{ij}$	(page C 29)
$\bar{a}_{ij}^{(k)}$	(page C 31)

Techniques

1. Show that the computed solution of $A\mathbf{x} = \mathbf{b}$ is the exact solution $\overline{\mathbf{x}}$ of the perturbed equations $(A + \delta A)\overline{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$ in which δA and $\delta \mathbf{b}$ can be large. * *
2. Show that Gauss elimination with interchanges (pivoting) gives small perturbations δA and $\delta \mathbf{b}$ and therefore exhibits very little induced instability. * *

8.4 ACCURACY AND ILL-CONDITIONING

8.4.1 Accuracy of Approximate Solution

The backward error analysis essentially gives an evaluation of our *technique*, but it does not give us an expression for the error in the computed solution of the linear equations. Any type of forward error analysis is virtually impossible, and in fact it is really rather difficult to get an accurate error bound for the solution.

If we substitute the alleged solution $\bar{\mathbf{x}}$ into the equations, and compute the *residual vector* $\mathbf{r} = \mathbf{b} - A\bar{\mathbf{x}}$, then the error in the solution is exactly $A^{-1}\mathbf{r}$, that is $\mathbf{x} = \bar{\mathbf{x}} + A^{-1}\mathbf{r}$. But the computation of A^{-1} is at least twice as laborious as the computation of $\bar{\mathbf{x}}$, and we would prefer to get some idea of the error without finding even an approximation to A^{-1}.

If the problem is at all ill-conditioned, then even small elements in the residual vector $\mathbf{r}$ do not guarantee that $\bar{\mathbf{x}}$ is a good approximation, since, as we have seen, $A^{-1}\mathbf{r}$ could have larger components than $\bar{\mathbf{x}}$ even if those of $\mathbf{r}$ are quite small. It can be proved, moreover, that if $\bar{\mathbf{x}}$ is small then $\mathbf{r}$ is certainly small if we use our stable version of Gauss elimination.

The only immediate information that the elements of the residual vector can provide is that we have not committed any serious blunders. On the other hand, in the process of computing $\bar{\mathbf{x}}$ we have effectively performed the approximate triangular decomposition of A, and we can, therefore, without much extra labour, solve the equations $A\delta\mathbf{x} = \mathbf{r}$ approximately to obtain a correction $\delta\mathbf{x} = A^{-1}\mathbf{r}$ to $\bar{\mathbf{x}}$.

Consider, for example, the equations

$$\begin{aligned}
0.5000x_1 + 0.3333x_2 + 0.2500x_3 &= 0.9500\\
0.2500x_1 + 0.2000x_2 + 0.1667x_3 &= 0.5200\\
0.3333x_1 + 0.2500x_2 + 0.2000x_3 &= 0.6700
\end{aligned}$$

ordered so that no interchanges are needed in the elimination or triangular decomposition. The solution, correctly rounded to four decimals, is

$$x_1 = 1.1887, \qquad x_2 = 0.6440, \qquad x_3 = 0.5641$$

Now suppose we work with four-figure floating-point arithmetic. First we find the triangular decomposition, which turns out to be

L			U		
1	0	0	0.5000	0.3333	0.2500
0.5000	1	0	0	0.03340	0.04170
0.6666	0.8323	1	0	0	−0.001310

Next we find $\bar{\mathbf{x}}$ from the equations $L\mathbf{c} = \mathbf{b}$, $U\bar{\mathbf{x}} = \mathbf{c}$, which give

$$\begin{aligned}
\mathbf{c} &= (0.9500,\ 0.045\,00,\ -0.000\,800\,0)\\
\bar{\mathbf{x}} &= (1.205,\ 0.5847,\ 0.6107)
\end{aligned}$$

and the approximate solution differs from the truth by a relative maximum of about 10%.

Let us try to improve this approximation. We first compute *exactly* the residual vector

$$\mathbf{r} = \mathbf{b} - A\bar{\mathbf{x}} = (-0.000\,055\,51,\ 0.000\,006\,31,\ 0.000\,058\,50).$$

We then round the elements of the residual vector to single-length floating-point form:

$$\bar{\mathbf{r}} = (10^{-4} \times -0.5551,\quad 10^{-5} \times 0.6310,\quad 10^{-4} \times 0.5850).$$

The rounding error is usually small, and here in fact zero, because the residual elements are themselves small, with several zeros after the decimal point.

We now correct the solution, solving in obvious notation and with single-length arithmetic the equations $L\delta \mathbf{c} = \bar{\mathbf{r}}$, $U\delta\mathbf{x} = \delta\mathbf{c}$, which give

$$\delta\mathbf{c} = 10^{-4}(-0.5551, \quad 0.3407, \quad 0.6714)$$
$$\delta\mathbf{x} = (-10^{-1} \times 0.1761, \quad 10^{-1} \times 0.6500, \quad -10^{-1} \times 0.5125)$$

The rounded better solution $\bar{\mathbf{x}} + \delta\mathbf{x}$ is

$$x_1 = 1.187, \quad x_2 = 0.6497, \quad x_3 = 0.5595,$$

which now differs from the truth by a relative maximum of about 1%, which is a full decimal figure better than the first approximation.

We can continue this process, first finding the exact new residual vector

$$\mathbf{r} = (0.000\,079\,99, \quad 0.000\,041\,35, \quad 0.000\,047\,90)$$

and a new correction $\delta\mathbf{x}$. This we do not pursue. We note, however, the interesting fact that the new residual elements, for this much better approximation, are no smaller than those of the poor first approximation! This confirms the inadequacy of the residual vector as an indication of the accuracy of the approximate solution.

It is important to realize that the process succeeds only if the residual elements are computed very accurately and if the resulting storage error is small. For otherwise the latter error, premultiplied by A^{-1}, could well be as large as our correction!

It may be argued that this accurate computation of residual elements involves the use of "double-length" computer arithmetic, suggesting that the complete solution could be obtained quite accurately by using double-length arithmetic right from the start. The point is that double-length arithmetic is expensive in storage and in time. The computation of the residual vector involves only n^2 numerical operations, considerably fewer than the approximately $\frac{1}{3}n^3$ operations involved in elimination and back-substitution for the large values of n which we meet in practice. We shall also see in sub-section 8.4.2 that this device has a further important benefit.

We can go even further with this idea. If the problem is mathematical we have to use rounded data in the solution of the linear equations, but we might be able to compute the residual vector using a more accurate or even exact version of A. The correction could then be obtained as before.

For instance, our stored matrix in the last example may be the rounded version of the true matrix

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}.$$

We could obtain the first approximate solution as before, but the correctly-computed and rounded residual vector of our approximate solution is now

$$\mathbf{r} = 10^{-4}(-0.7500, \quad 0.2667, \quad 0.1833).$$

The correction vector obtained from this residual vector is

$$\delta\mathbf{x} = (-0.005\,212, \quad 0.016\,13, \quad -0.011\,38)$$

and the rounded corrected solution is therefore

$$\mathbf{x} = (1.200, \quad 0.6008, \quad 0.5993).$$

This has a maximum error of 0.0008 compared with the exact solution

$$\mathbf{x} = (1.200, \quad 0.6000, \quad 0.6000),$$

whereas the first approximation has a maximum error of 0.0162.

In all cases we can repeat the whole operation and expect to get even better answers.

8.4.2 Practical Determination of Ill-conditioning

The method of correcting a first approximation which we have just described is also usually the most powerful method available for detecting the possibility of ill-conditioning. For, as we have seen in the error analysis of the Gauss elimination method, our first approximation $\bar{\mathbf{x}}$ is the exact solution for a perturbed problem $(A + \delta A)\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$, where the perturbations δA and $\delta\mathbf{b}$ have known upper bounds. The corrected approximation is effectively the exact solution of the original problem with smaller perturbations, and the number of figures to which our two approximations agree therefore gives useful information about the degree of ill-conditioning.

In an ill-conditioned mathematical problem, in which the (storage) uncertainty in the original data is at least no greater than the perturbation induced by the method, examination of successive approximate solutions gives not only an idea about the ill-conditioning but also a reasonable guarantee of the number of figures we can quote as being correct in the solution. There may, of course, be some limit to this, because if we have to round the elements of the residual vector, there is an unavoidable error $A^{-1}\delta\mathbf{r}$. Even though $\delta\mathbf{r}$ is of order 2^{-2t} the elements of $A^{-1}\delta\mathbf{r}$ might, in extreme cases, affect some of even the first t digits of the solution.

For a physical problem, the real uncertainties in the data are, as we have seen, likely to be very much larger than the induced perturbations in the method. In that case we have a virtual guarantee that the number of *meaningful* figures in the result is not greater than the number of figures to which our first and second approximations agree. It may be very much smaller, but we have no easy way of deciding this.

What we would really like to do is to find accurate intervals in which the x_r certainly lie, corresponding to the different data intervals in which the a_{rs} and b_r lie. This is an extremely difficult problem, and the subject of much current research. Nevertheless the partial solution of the last paragraph is already an important and useful result.

8.4.3 Summary of Section 8.4

In this section we defined the term

residual vector (page C 36) * *

and introduced the notation for it

$\mathbf{r}$ (page C 36)

Techniques

1. Compute the residual vector for an alleged solution $\bar{\mathbf{x}}$ of $A\mathbf{x} = \mathbf{b}$ and hence determine $\delta\mathbf{x}$, a first correction to $\bar{\mathbf{x}}$. * *
2. Use the size of $\delta\mathbf{x}$ to determine the degree of ill-conditioning of the problem. * *

8.5 SUMMARY OF THE UNIT

In this unit we have discussed and illustrated the following items.

1. The fact that the solution of sets of linear algebraic equations can be very sensitive to small changes in the data, the coefficients of the matrix or the constant terms in the equations. In other words, that this problem can exhibit ill-conditioning.
2. The uncertainties in the answers may be correlated, so that, for example, an allied problem such as the evaluation of some linear combination of the elements of the solution vector, may be quite well-conditioned.
3. The elimination methods of Gauss and Jordan, and their relation with the production of the Hermite normal form, for solving systems of equations or inverting matrices.
4. The connection between the reduction of a matrix to upper triangular form, by elimination, and the expression of a matrix as the product of lower and upper triangular matrices.
5. The effect of zero diagonal elements in the elimination method, the corresponding need for row interchanges, and the connection between this phenomenon and the singularity of leading principal sub-matrices.
6. The fact that the elimination method can produce very poor results because the method can exhibit induced instability.
7. An error analysis, relating the induced instability to the fact that the computed solution is the exact solution of a perturbed problem, in which the perturbations in the matrix and right-hand-vector can be very large.
8. A method of avoiding the induced instability, giving rise to guaranteed stability, which merely involves selected row interchanges so that the largest element in the relevant column is on the diagonal. This is "Gauss elimination with interchanges, or with pivoting", and the error analysis gives a guaranteed small bound for the induced perturbations.
9. That the error analysis involved, so-called "backward" error analysis, gives an evaluation of the method but not a bound for the accuracy of the computed solution.
10. A method for correcting the approximate solution, which involves little extra work and which also exhibits the degree of ill-conditioning of the problem.

Definitions

The terms defined in this unit and page references to their definitions are given below

direct method	(page C 6)	* * *
iterative method	(page C 6)	* * *
physical problem	(page C 7)	* *
mathematical problem	(page C 9)	* *
Gauss elimination method	(page C 12)	* * *
upper triangular matrix	(page C 12)	* * *
back-substitution	(page C 12)	* * *
Jordan method	(page C 17)	* * *
pivot	(page C 18)	* * *
essential row interchanges	(page C 20)	* *
unit lower triangular matrix	(page C 21)	* *
triangular decomposition of a matrix	(page C 21)	* *
compact elimination method	(page C 25)	*
forward error analysis	(page C 28)	* *
backward error analysis	(page C 28)	* *
pivoting	(page C 32)	* *
residual vector	(page C 36)	* *

Techniques

1.	Solve $A\mathbf{x} = \mathbf{b}$, given specified errors in A or $\mathbf{b}$ and hence say whether or not the problem is ill-conditioned.	*
2.	If such a problem is ill-conditioned determine whether the problem of finding $\sum_{s=1}^{n} c_s x_s$, given $c_1, c_2, \ldots, c_n$, is well-conditioned.	*
3.	Solve $A\mathbf{x} = \mathbf{b}$ and invert A by the Gauss elimination method and by the Jordan method.	*
4.	Determine the triangular decomposition of A.	* * *
5.	Solve $A\mathbf{x} = \mathbf{b}$ by the compact elimination method.	* * *
6.	Evaluate det A by reducing A to triangular form.	* * *
7.	Show that the computed solution of $A\mathbf{x} = \mathbf{b}$ is the exact solution $\bar{\mathbf{x}}$ of the perturbed equations $(A + \delta A)\bar{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$ in which δA and $\delta\mathbf{b}$ can be large.	* * *
8.	Show that Gauss elimination with interchanges (pivoting) gives small perturbations δA and $\delta\mathbf{b}$ and therefore exhibits very little induced instability.	* *
9.	Compute the residual vector for an alleged solution $\bar{\mathbf{x}}$ of $A\mathbf{x} = \mathbf{b}$ and hence determine $\delta\mathbf{x}$, a first correction to $\bar{\mathbf{x}}$.	* *
10.	Use the size of $\delta\mathbf{x}$ to determine the degree of ill-conditioning of the problem.	* *

Notation

m_{ij}	(page C 12)
$\mathbf{b}^{(k)}$	(page C 12)
$A^{(k)}$	(page C 12)
U	(page C 12)
L	(page C 21)
$L(r)$	(page C 22)
$U(s)$	(page C 22)
δA	(page C 28)
$\delta\mathbf{b}$	(page C 28)
$\overline{m}_{ij}$	(page C 29)
$\bar{a}_{ij}^{(k)}$	(page C 31)
$\mathbf{r}$	(page C 36)

8.6 SELF-ASSESSMENT

Self-assessment Test

This Self-assessment Test is designed to help you test quickly your understanding of the unit. It can also be used, together with the summary of the unit for revision. The answers to these questions will be found on the next non-facing page. We suggest you complete the whole test before looking at the answers.

1. Find, by the method of *Jordan* elimination, the inverse of each of the matrices

$$A_1 = \begin{bmatrix} 1 & 0.9 \\ 1 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 1 \\ 1 & -3 \end{bmatrix}.$$

2. We wish to solve the equations $A\mathbf{x} = \mathbf{b} + \mathbf{e}$, where $\mathbf{e}$ is a vector of uncertainties in the measured vector $\mathbf{b}$, and each element of $\mathbf{e}$ is at most ε in absolute value.

   For which of the matrices of Question 1 is the problem absolutely ill-conditioned, in the sense that the uncertainties in the components of the solution vector $\mathbf{x}$ are greater than those of the uncertainty vector $\mathbf{e}$?

3. Which of the following problems are (a) absolutely ill-conditioned, and (b) relatively ill-conditioned, in the sense that the relative uncertainties in the components of the solution are greater than the relative uncertainties in the vector $\mathbf{b}$?

   (i) $x_1 + 0.9x_2 = 1 + e_1$
   $x_1 + \quad x_2 = -1 + e_2$

   (ii) $x_1 + 0.9x_2 = 1 + e_1$
   $x_1 + \quad x_2 = 1 + e_2$

   where $|e_1|, |e_2| \leqslant 10^{-3}$.

4. In the problem (ii) of Question 3, show that all we can say about x_1 and x_2 is

$$x_1 = 1 \pm 19 \times 10^{-3} \qquad x_2 = 0 \mp 20 \times 10^{-3}$$

   but that, for the computation of $y = x_1 + x_2$, we can assert with confidence that

$$y = 1 \pm 10^{-3}.$$

5. We wish to solve linear equations $A\mathbf{x} = \mathbf{b}$, with

$$A = \begin{bmatrix} 1 & 0.9 & 1 \\ 1 & 1 & 2 \\ -1 & 0.1 & 3 \end{bmatrix}$$

   Is pivoting necessary in Gauss elimination to avoid induced instability? If it is, write down a matrix $\bar{A}$, a row-permutation of A, for which pivoting is not necessary.

6. From your work in Question 5, write down the determinant of the matrix A of that question.

7. Carry out triangular decompositions of (i) the matrix A_1 in Question 1, and (ii) the row-permuted $\bar{A}$ of Question 5.

8. Working with one-digit floating-point arithmetic, the Gauss elimination process for solving the equations

$$\begin{aligned} 0.8x_1 + 0.6x_2 &= 0.2 \\ -0.5x_1 + 0.4x_2 &= -0.8 \end{aligned}$$

produces, in the process of elimination, the arrays

multipliers		*A*	**b**
	0.8	0.6	0.2
0.6	−0.5	0.6	0.2

$A^{(2)}$		$\mathbf{b}^{(2)}$
0.8	0.6	0.2
0	0.8	−0.7

Write down a matrix $A + \delta A$ and a vector $\mathbf{b} + \delta\mathbf{b}$ which, with exact arithmetic, would produce exactly the *recorded* multipliers, the recorded upper triangle $A^{(2)}$, and the *recorded* vector $\mathbf{b}^{(2)}$ in this elimination.

9. With the same floating-point arithmetic, the back substitution applied in Question 8 leads to the approximate solution

$x_1 = 0.9$, $x_2 = -0.9$.

Show that the true solution of the equation is $x_1 + \delta x_1$, $x_2 + \delta x_2$, where

$$\begin{bmatrix} 0.8 & 0.6 \\ -0.5 & 0.4 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = \begin{bmatrix} 0.02 \\ 0.01 \end{bmatrix}$$

10. Compute the correction, $\delta\mathrm{x}$ of Question 9, working with one-digit floating-point arithmetic, and verify that the corrected value is a better approximation to the true solution (which is $x_1 = 0.903$, $x_2 = -0.871$, correct to three decimal places).

Solution to Self-assessment Test

1.

$$\begin{array}{rrrrr} & & A_1 & & \\ & 1 & 0.9 & 1 & 0 \\ -1 & 1 & 1 & 0 & 1 \\ & & & & \\ -9 & 1 & 0.9 & 1 & 0 \\ & 0 & 0.1 & -1 & 1 \\ & & & & \\ & 1 & 0 & 10 & -9 \\ & 0 & 0.1 & -1 & 1 \\ & \multicolumn{2}{c}{I} & \multicolumn{2}{c}{A_1^{-1}} \\ & 1 & 0 & 10 & -9 \\ & 0 & 1 & -10 & 10 \end{array} \qquad \begin{array}{rrrrr} & & A_2 & & \\ & 1 & 1 & 1 & 0 \\ -1 & 1 & -3 & 0 & 1 \\ & & & & \\ \frac{1}{4} & 1 & 1 & 1 & 0 \\ & 0 & -4 & -1 & 1 \\ & & & & \\ & 1 & 0 & \frac{3}{4} & \frac{1}{4} \\ & 0 & -4 & -1 & 1 \\ & \multicolumn{2}{c}{I} & \multicolumn{2}{c}{A_2^{-1}} \\ & 1 & 0 & \frac{3}{4} & \frac{1}{4} \\ & 0 & 1 & \frac{1}{4} & -\frac{1}{4} \end{array}$$

2. The uncertainty in the solution is $\delta\mathbf{x} = A^{-1}\mathbf{e}$. For A_1, we have (using the inverse in Question 1)

$$\delta x_1 = 10e_1 - 9e_2$$
$$\delta x_2 = -10e_1 + 10e_2 .$$

If $|e_1|$, $|e_2| \leqslant e$, then the maxima of $|\delta x_1|$, $|\delta x_2|$ are $19e$ and $20e$. These are much greater than $|e_1|$ and $|e_2|$, and the problem is absolutely ill-conditioned.

For A_2, we have

$$\delta x_1 = \tfrac{3}{4}e_1 + \tfrac{1}{4}e_2$$
$$\delta x_2 = \tfrac{1}{4}e_1 - \tfrac{1}{4}e_2 .$$

The maxima of $|\delta x_1|$, $|\delta x_2|$ are e and $\frac{1}{2}e$, and the problem is absolutely well-conditioned.

3. The matrix is A_1 of Question 1. Using its inverse, we find, for (i),

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10 & -9 \\ -10 & 10 \end{bmatrix} \begin{bmatrix} 1 + e_1 \\ -1 + e_2 \end{bmatrix} = \begin{bmatrix} 19 \\ -20 \end{bmatrix} + \begin{bmatrix} 10 & -9 \\ -10 & 10 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

The maximum values of the uncertainties are 19×10^{-3} and 20×10^{-3} in absolute value. But these are only 1 in 10^3 of the values of x_1 and x_2, so that the problem, though absolutely ill-conditioned, is relatively well-conditioned.

For (ii), we find

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \text{uncertainties.}$$

The maximum uncertainties are the same as before, but they are large relative to the value of x_2, and the problem is ill-conditioned, both relatively and absolutely.

4. The maximum uncertainties in part (ii) of Question 3 occur when $e_1 = \pm 10^{-3}$, $e_2 = \pm 10^{-3}$ and all we can then say is

$$x_1 = 1 \pm 19 \times 10^{-3}$$
$$x_2 = 0 \mp 20 \times 10^{-3}.$$

The value of

$$\begin{aligned} x_1 + x_2 &= [1 \quad 1] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= [1 \quad 1] \begin{bmatrix} 1 + 10e_1 - 9e_2 \\ 0 - 10e_1 + 10e_2 \end{bmatrix} \\ &= 1 + (10e_1 - 9e_2) + (-10e_1 + 10e_2) \\ &= 1 + e_2 , \end{aligned}$$

and since $|e_2| \leqslant 10^{-3}$, we can say that $y = 1 \pm 10^{-3}$, a reasonably well-conditioned situation.

5. The elimination goes as follows, with pivots emboldened.

$$\begin{array}{lccccl} \textit{multipliers} & & A & & & \\ & \mathbf{1} & 0.9 & 1 & & \text{(No interchange needed here.)} \\ -1 & 1 & 1 & 2 & & \\ 1 & -1 & 0.1 & 3 & & \end{array}$$

$$\begin{array}{lccccl} & & A^{(2)} & & & \\ & 1 & 0.9 & 1 & & \text{(There is a larger element in} \\ & 0 & 0.1 & 1 & & \text{column 2 below the diagonal,} \\ & 0 & 1.0 & 4 & & \text{interchange needed.)} \end{array}$$

Interchange rows 2 and 3.

$$\begin{array}{lccccl} \textit{multipliers} & & \bar{A}^{(2)} & & & \\ & 1 & 0.9 & 1 & & \\ & 0 & \mathbf{1.0} & 4 & & \\ -0.1 & 0 & 0.1 & 1 & & \\ & 1 & 0.9 & 1 & & \\ & 0 & 1.0 & 4 & & \text{(Final upper triangular form.)} \\ & 0 & 0 & 0.6 & & \end{array}$$

Pivoting is therefore not necessary if we interchange rows 2 and 3 before starting the elimination, that is, starting with

$$\bar{A} = \begin{bmatrix} 1 & 0.9 & 1 \\ -1 & 0.1 & 3 \\ 1 & 1 & 2 \end{bmatrix}.$$

6. The determinant of $\bar{A}$ in Question 5 is the product of the diagonal terms of the final upper triangular form, that is

$$\det \bar{A} = 0.6.$$

$\bar{A}$ is obtained from A with one row interchange, so

$$\det A = -0.6.$$

7.

$$A_1 = \begin{bmatrix} 1 & 0.9 \\ 1 & 1 \end{bmatrix} = \overset{L}{\begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix}} \overset{U}{\begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}}$$

Then, using a_{rs} = product of rth row of L with sth column of u, we find

$$\begin{array}{ll} (1,1) & 1 = u_{11} \\ (1,2) & 0.9 = u_{12} \\ (2,1) & 1 = l_{21}\, u_{11},\ \text{giving } l_{21} = 1 \\ (2,2) & 1 = l_{21}\, u_{12} + u_{22},\ \text{giving } u_{22} = 1 - 0.9 = 0.1. \end{array}$$

We can use the same method for the matrix $\bar{A}$ of Question 5. But it is easier to use the relation between triangular decomposition and elimination (which we have already performed). This is

$$\bar{A} = \begin{bmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & -m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

Since the rows 2 and 3 have been interchanged in the elimination involving A, however, we must interchange the multipliers $-m_{21}$ and $-m_{31}$, so that they refer to their "proper" rows. This gives

$$\bar{A} = \overset{L}{\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0.1 & 1 \end{bmatrix}} \overset{U}{\begin{bmatrix} 1 & 0.9 & 1 \\ 0 & 1.0 & 4 \\ 0 & 0 & 0.6 \end{bmatrix}},$$

which can be verified by direct matrix multiplication.

8. The easiest way to find the perturbed matrix $A + \delta A$ is to find the matrix for which the *stored* L (obtained from the *stored* multipliers) and the *stored* U (the final *stored* upper triangle) is the exact triangular decomposition. That is,

$$A + \delta A = LU = \begin{bmatrix} 1 & 0 \\ -0.6 & 1 \end{bmatrix} \begin{bmatrix} 0.8 & 0.6 \\ 0 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.6 \\ -0.48 & 0.44 \end{bmatrix}.$$

Similarly,

$$\mathbf{b} + \delta\mathbf{b} = L\mathbf{b}^{(2)} = \begin{bmatrix} 1 & 0 \\ -0.6 & 1 \end{bmatrix} \begin{bmatrix} 0.2 \\ -0.7 \end{bmatrix} = \begin{bmatrix} 0.2 \\ -0.82 \end{bmatrix}.$$

The elimination applied to

$$\begin{aligned} 0.8x_1 \ + 0.6x_2 &= 0.2 \\ -0.48x_1 + 0.44x_2 &= -0.82 \end{aligned}$$

can then be performed *exactly* with one-figure floating-point decimal arithmetic.

9. If $\bar{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, the computed first approximation, we find the residual vector

$$\mathbf{r} = \mathbf{b} - A\bar{\mathbf{x}}$$

The correct solution satisfies $A\mathbf{x} = \mathbf{b}$, so that the correction $\delta\mathbf{x} = \mathbf{x} - \bar{\mathbf{x}}$ satisfies exactly the equation

$$\mathbf{r} = A\mathbf{x} - A\bar{\mathbf{x}} = A(\mathbf{x} - \bar{\mathbf{x}}) = A\delta\mathbf{x}.$$

We find

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} 0.2 \\ -0.8 \end{bmatrix} - \begin{bmatrix} 0.8 & 0.6 \\ -0.5 & 0.4 \end{bmatrix} \begin{bmatrix} 0.9 \\ -0.9 \end{bmatrix} = \begin{bmatrix} 0.02 \\ 0.01 \end{bmatrix},$$

and the equations given in the question are $A\delta\mathbf{x} = \mathbf{r}$.

10. For the solution of the equations for the correction $\delta\mathbf{x}$, which we can only get approximately with the use of one-digit arithmetic, we use the same elimination method, noting that the "elimination in the matrix" has already been performed. We have

(i) *elimination*

	A		$\mathbf{r}$
	0.8	0.6	0.02
0.6	−0.5	0.4	0.01
	0.8	0.6	0.02
	0	0.8	0.02

(ii) *back-substitution*

$$\delta x_2 = \text{fl}\,(0.02/0.8) = 0.02, \text{ in our arithmetic.}$$

$$\delta x_1 = \text{fl}\left(\frac{0.02 - 0.6(0.02)}{0.8}\right) = 0.01.$$

The hoped-for better approximation is $x_1 = 0.91$, $x_2 = -0.88$. At least the maximum error has been reduced from 0.029 to 0.09. (But the roundings here are quite marginal. We could equally accurately take $\delta x_2 = 0.02/0.8 = 0.03$ in our arithmetic; and then $\delta x_1 = 0$. The corrected solution is then $x_1 = 0.90$, $x_2 = 0.87$, with errors of only 0.003 and 0.001 respectively.)

LINEAR MATHEMATICS

1 Vector Spaces
2 Linear Transformations
3 Hermite Normal Form
4 Differential Equations I
5 Determinants and Eigenvalues
6 NO TEXT
7 Introduction to Numerical Mathematics: Recurrence Relations
8 Numerical Solution of Simultaneous Algebraic Equations
9 Differential Equations II: Homogeneous Equations
10 Jordan Normal Form
11 Differential Equations III: Nonhomogeneous Equations
12 Linear Functionals and Duality
13 Systems of Differential Equations
14 Bilinear and Quadratic Forms
15 Affine Geometry and Convex Cones
16 Euclidean Spaces I: Inner Products
17 NO TEXT
18 Linear Programming
19 Least-squares Approximation
20 Euclidean Spaces II: Convergence and Bases
21 Numerical Solution of Differential Equations
22 Fourier Series
23 The Wave Equation
24 Orthogonal and Symmetric Transformations
25 Boundary-value Problems
26 NO TEXT
27 Chebyshev Approximation
28 Theory of Games
29 Laplace Transforms
30 Numerical Solution of Eigenvalue Problems
31 Fourier Transforms
32 The Heat Conduction Equation
33 Existence and Uniqueness Theorem for Differential Equations
34 NO TEXT

335 01097 0